预测模型与 ROC 曲线

本模块用于构建预测模型，如果结果变量为两分类变量或时间依赖的生存状态变量，自动进行 ROC 分析。可以设置建模人群与验证人群，一种方法是随机按比例把所有观察样本分成 Development set （D set）与 Validation set (V set)，第二种方法通过变量赋值定义构建模型所用样本如：FMYTYPE=0,表示数据中 FMYTYPE=0 的样本用来构建模型（D set），其它的用来验证模型(V set)。

构建预测模型的流程：

1）首先对所列自变量进行共线性筛查（VIF），剔除 VIF>10 的变量。
2）检查自变量中是否有连续型变量，如果某 X 是连续性变量，它与结果变量函数 f(Y)的关系不一定是直线性的，本模块自动采用多变量分数多项式（Multivariable Fractional Polynomials）方法确定每个连续型变量是否要添加多次项，即 X^2 (平方),X^3（三次方）,X^0.5（平方根）。并自动筛除不必要的自变量，输出 MFP 方法得出的模型。参考文献：https://cran.r-project.org/web/packages/mfp/vignettes/mfp_vignette.pdf
3）建立全自变量的预测模型（full model）。
4）运用逐步回归筛选自变量建立简化的模型(stepwise model)。
5）如果定义了 bootstrap 次数，采用重抽样的方法，产生 n 个新样本，分别构建回归模型，计算每个自变量的回归系数的均值。即 BS full model 与 BS stepwise model。
6）对上述 5 种模型进行验证比较，帮助用户选择最优模型。

一、对两分类结果变量的预测模型

对两分类的结果变量，采用 logistic 回归模型进行预测。由预测模型计算得出的预测分数是介于 0 至 1 之间的连续性值，对这个分数取切点用来判断结果发生与否。验证预测效果常用的是 ROC 曲线及该曲线下的面积（AUC），常用指标有：

1) 敏感性：敏感性就是指由金标准确诊有病组内所检测出阳性病例数的比率（%）。即本实验诊断的真阳性率。其敏感性越高，漏诊的机会就越少。

2) 特异性：是指由金标准确诊为无病组内所检测出阴性人数的比率（%），即本诊断实验的真阴性率。特异性越高，发生误诊的机会就越少。

3) 诊断准确率：是指临床诊断检测出的真阳性和真阴性例数之和，占总检测人数的比例，即称本临床实验诊断的准确性。

4) 阳性似然比（positive likelihood ratio）：　阳性似然比是指临床诊断检测出的真阳性率与假阳性率之间的比值，即阳性似然比=敏感性/（1-特异性）。可用以描述诊断试验阳性时，患病与不患病的机会比。提示正确判断为阳性的可能性是错误判断为阳性的可能性的倍数。阳性似然比数值越大，提示能够确诊患有该病的可能性越大。它不受患病率影响，比起敏感度和特异度更为稳定。

5) 阴性似然比（negative liklihoodratio）：阴性似然比是指临床实验诊断检测出的假阴性率与真阴性率之比值，此值越小，说明该诊断方法越好。可用以描述诊断试验阴性

时，患病与不患病的机会比。阴性似然比提示错误判断为阴性的可能性是正确判断为阴性的可能性的倍数。阴性似然比数值越小，提示能够否定患有该病的可能性越大。

6) Yuden 指数: Yuden 指数 = 敏感性+特异性-1

7) ROC 曲线：称受试者工作曲线，可以综合考虑一项诊断试验(定量指标)或预测模型（模型的预测值）在所有诊断界值时的灵敏度和特异度。对于每一个诊断界值，都可以得到相应的灵敏度和特异度。ROC 曲线是以（1-特异度）为横坐标，以灵敏度为纵坐标绘制而成的曲线，它用线段连接每个诊断界值对应的[（1-特异度），灵敏度]的点。由预测模型计算出来的是一个连续性的对结果 Y 的预测值。对于这种连续变量，诊断界值可以取任意一个。对有序分类变量，由不同的诊断结果作为诊断界值时，对应于不同的灵敏度和特异度，将每种诊断结果对应的[（1-特异度），灵敏度]的点，标在直角坐标系中，用线段连接各相邻两点，即为有序分类资料的 ROC 曲线。

8) ROC 曲线下面积（AUC）：表示所有灵敏度时诊断试验平均 特异度，或者所有特异度时诊断试验的平均灵敏度。通常，ROC 曲线下面积在 0.5-1 之间。曲线下从原点到右上角的对角线称为机会线，表示无论取任诊断界值，灵敏度=1-特异度，即真阳性率=假阳性率，意味着无论患者和非患者都有同样的“机会”被诊断为阳性。ROC 曲线越接近机会线，即曲线下面积越接近 0.5，表明诊断试验区分患者和非患者的能力越弱；越接近 1，表明诊断试验的准确度越强。一般认为，0.50-0.70 之间，诊断价值较小；0.70-0.90 之间，诊断价值中等；>0.90，诊断价值较高。最理想的诊断试验的 ROC 曲线是从坐标原点出发，沿着 Y 轴到（0，1）点，再沿着 X 轴的水平线到（1，1）点。在比较 ROC 曲线下面积时，还应考虑到实际临床应用情况。比如，某项诊断试验主要用于排除疾病时，则需要较高的特异度，这是我们仅对左侧的 ROC 曲线（即高特异度的 ROC 曲线部分）下的面积感兴趣。通过检验 AUC 是否等于 0.5 来评价某诊断试验有无诊断价值。

9) 诊断界值确定：实际工作中，人们希望找到灵敏度和特异度均接近“1”的点。横轴为（1-特异度），所以横轴原点就是特异度为 1 的点，因此我们要找的点就是距 ROC 曲线图中左上角最近的点，也就是（灵敏度+特异度）取最大值的点。如果认为灵敏度的重要性是特异度的 a 倍，此时可选取（a*灵敏度+1*特异度）取值最大的点。在实际应用中，可以根据不同的研究目的确定阈值，如果诊断试验目的是筛查本病时，宜选在误诊率充许的范围内灵敏度较高的截断点，此时保证了漏诊率低；若试验目的为确诊本病，则宜选在漏诊率充许范围内特异度较高的截断点，此时误诊率低。

二、对连续性变量 Y 的预测模型

如果 Y 是连续性变量，采用线性拟合构建模型（要求残差接近正态分布），该模型得出 Y 的预测值。将预测值与原观察值进行一致性分析，以判断预测模型的优劣。一致性分析方法采用 Bland-Altman 方法，具体参考“定量测量方法比较”。

三、对时间依赖的生存状态变量的预测模型（cox 模型）

如果 Y 是生存状态变量（0=censored 1=发生），构建 COX 回归模型。模型构建后，采用 R timeROC 包，用 Inverse Probability of Censoring Weighting (IPCW) 估计累计/动态时间依赖的 ROC 曲线。

参考文献：

Hung, H. and Chiang, C. (2010). Estimation methods for time-dependent AUC with survival data. Canadian Journal of Statistics, 38(1):8-26

Uno, H., Cai, T., Tian, L. and Wei, L. (2007). Evaluating prediction rules for t-years survivors with censored regression models. Journal of the American Statistical Association, 102(478):527-537.

实例分析

例 1：练习项目 DEMO 构建 HBP(是否高血压：0=否 1=是)的预测模型，使用 FMYTYPE=1 的样本构建模型，其它用作模型验证。输入界面如。

输出结果：

预测模型与 ROC 分析

```
Outcome: High BP
Collinearity VIF selection:
```

	Step 1
AGE	1.4

HEIGHT	3
WEIGHT	1.8
OCCU.NEW	1.1
EDU.NEW	1.8
SEX	2.4

Variables removed:
Variables selected: AGE HEIGHT WEIGHT OCCU.NEW EDU.NEW SEX

Models:

Model 0: Multiple Fractional Polynomial model from observed data
-3.81625 +7.53963*(AGE/100) -0.60500*(OCCU.NEW=2) +0.15819*(EDU.NEW=2) -0.39906*(EDU.NEW=3)

	Estimate	Std error	OR	95%CI.low	95%CI.upp	P-value
Intercept	-3.8163	0.4056	0.0220	0.0099	0.0487	0.0000
AGE/100	7.5396	0.7715	1881.1399	414.6717	8533.7074	0.0000
factor(OCCU.NEW)2	-0.6050	0.2009	0.5461	0.3683	0.8096	0.0026
factor(EDU.NEW)2	0.1582	0.2318	1.1714	0.7437	1.8451	0.4950
factor(EDU.NEW)3	-0.3991	0.2888	0.6710	0.3809	1.1819	0.1671

Model 1: Full model from observed data
1.97000 +0.07682*AGE -5.09560*HEIGHT +0.04634*WEIGHT -0.64375*(OCCU.NEW=2) +0.11859*(EDU.NEW=2) -0.46777*(EDU.NEW=3) -0.43839*(SEX=2)

	Estimate	Std error	OR	95%CI.low	95%CI.upp	P-value
(Intercept)	1.9700	3.3358	7.1707	0.0104	4954.6831	0.5548
AGE	0.0768	0.0085	1.0798	1.0620	1.0979	0.0000
HEIGHT	-5.0956	2.2827	0.0061	1e-040	0.5371	0.0256
WEIGHT	0.0463	0.0172	1.0474	1.0126	1.0834	0.0072
factor(OCCU.NEW)2	-0.6437	0.2087	0.5253	0.3490	0.7908	0.0020
factor(EDU.NEW)2	0.1186	0.2628	1.1259	0.6726	1.8847	0.6519
factor(EDU.NEW)3	-0.4678	0.3429	0.6264	0.3199	1.2267	0.1725
factor(SEX)2	-0.4384	0.3196	0.6451	0.3448	1.2070	0.1702

Model 2: Stepwise selected model from observed data
-0.83515 +0.08124*AGE -3.55888*HEIGHT +0.04395*WEIGHT -0.55739*(OCCU.NEW=2)

	Estimate	Std error	OR	95%CI.low	95%CI.upp	P-value
(Intercept)	-0.8352	2.1842	0.4338	0.0060	31.3708	0.7022
AGE	0.0812	0.0076	1.0846	1.0685	1.1010	0.0000
HEIGHT	-3.5589	1.6763	0.0285	0.0011	0.7609	0.0338
WEIGHT	0.0440	0.0172	1.0449	1.0104	1.0807	0.0104
factor(OCCU.NEW)2	-0.5574	0.1999	0.5727	0.3870	0.8474	0.0053

预测模型 ROC 曲线分析及最佳阈值分析(建模数据)

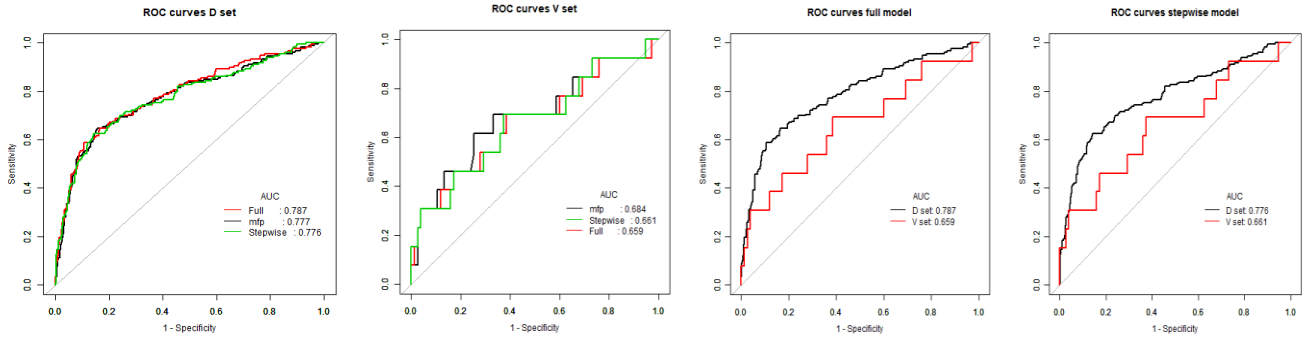
检测项目	mfp D set	Full D set	Stepwise D set
1	179	179	179
0	521	521	521
ROC 面积(AUC)	0.7771	0.7868	0.7764
95% 区间下限	0.7340	0.7452	0.7333
95% 区间上限	0.8202	0.8284	0.8194
最佳阈值	-0.6808	-0.7128	-0.6124
特异度	0.8426	0.8349	0.8580
敏感度	0.6480	0.6480	0.6257
准确度	0.7929	0.7871	0.7986
阳性似然比	4.1175	3.9259	4.4053
阴性似然比	0.4177	0.4215	0.4363
诊断比值比	9.8575	9.3134	10.0976
诊断需要检测数	2.0381	2.0705	2.0676
阳性预测值	0.5859	0.5743	0.6022
阴性预测值	0.8745	0.8735	0.8696
a	116	116	112
b	82	86	74
c	63	63	67
d	439	435	447

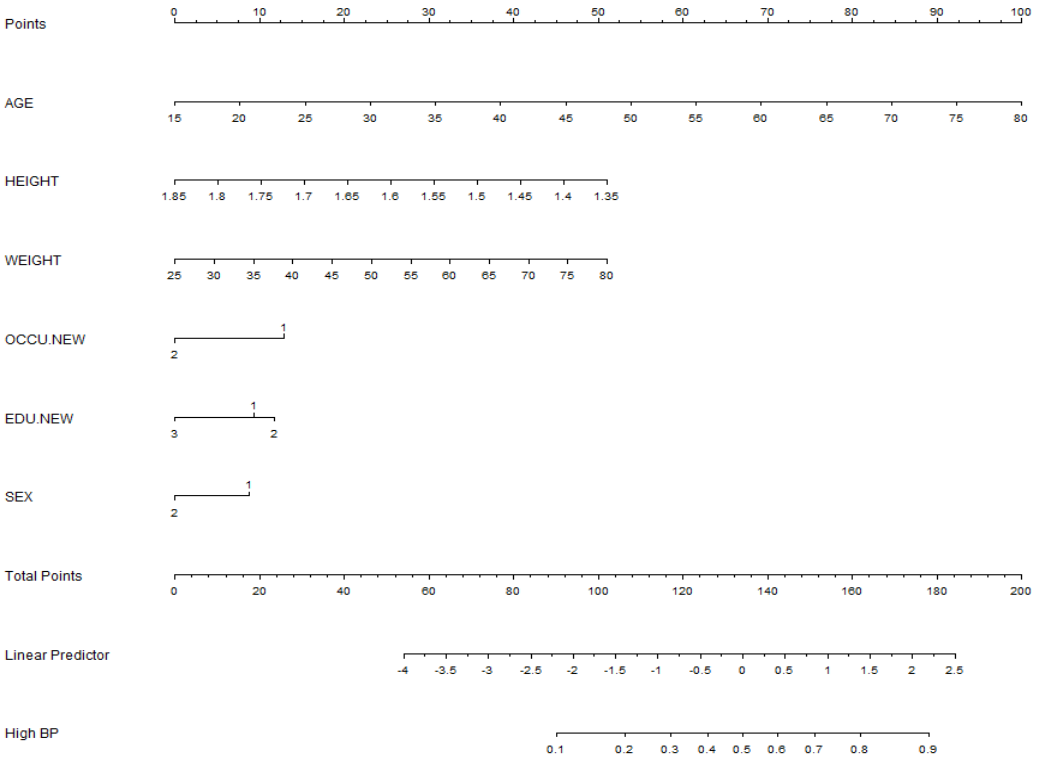
预测模型 ROC 曲线分析及最佳阈值分析(验证数据)

检测项目	mfp V set	Full V set	Stepwise V set
1	13	13	13
0	75	75	75
ROC 面积(AUC)	0.6836	0.6595	0.6605
95% 区间下限	0.5035	0.4757	0.4803
95% 区间上限	0.8636	0.8432	0.8407
最佳阈值	-0.8312	-1.1351	-1.1849
特异度	0.7467	0.6133	0.6267
敏感度	0.6154	0.6923	0.6923
准确度	0.7273	0.6250	0.6364
阳性似然比	2.4291	1.7905	1.8544
阴性似然比	0.5151	0.5017	0.4910
诊断比值比	4.7158	3.5690	3.7768
诊断需要检测数	2.7620	3.2718	3.1350
阳性预测值	0.2963	0.2368	0.2432

阴性预测值	0.9180	0.9200	0.9216
a	8	9	9
b	19	29	28
c	5	4	4
d	56	46	47

最佳阈值取敏感度+特异度最大的分界值。各分界点对应的敏感度特异保存在 ROC 输出文件（.xls）里





例 2：练习项目 DEMO 构建 SBP(收缩压)的预测模型，使用 FMYTYPE=1 的样本构建模型，其它用作模型验证。输入界面如。

输出结果：

预测模型与 ROC 分析

Outcome: Systolic BP, mmhg

Collinearity VIF selection:

	Step 1
AGE	1.4
HEIGHT	3
WEIGHT	1.8
OCCU.NEW	1.1
EDU.NEW	1.8
SEX	2.4

Variables removed:
Variables selected: AGE HEIGHT WEIGHT OCCU.NEW EDU.NEW SEX

Models:

Model 0: Multiple Fractional Polynomial model from observed data
323.60848 -124.77097*(AGE/100)^0.5 +221.00131*(AGE/100)^0.5 * log((AGE/100)) -
2.20601*(OCCU.NEW=2) +25.38172*(WEIGHT/100)

	Coeff.	Se.	95%CI.low	95%CI.upp	P.value
Intercept	323.6085	41.3378	242.5864	404.6306	<0.0001
(AGE/100)^0.5	-124.7710	36.1607	-195.6460	-53.8959	0.0006
AGE/100)^0.5 * log((AGE/100)	221.0013	34.6767	153.0350	288.9676	<0.0001
factor(OCCU.NEW)2	-2.2060	1.5669	-5.2771	0.8650	0.1596
WEIGHT/100	25.3817	9.9620	5.8563	44.9072	0.0111

Model 0: predicted vs. observed **(development data)**

Summary

Method	#obs	Minimum	Median	Maximum
Observed	700	92	125	255
Predicted	700	115.4585	125.614	176.02

Limits of agreement (assume slope=1)

Diff:(Predicted-Observed)	2.5% Limit	97.5% Limit	SD
0.00000	-38.85021	38.85021	19.42510

Correlation between predicted and observed: 0.5368

Model 0: predicted vs. observed **(validation data)**

Summary

Method	#obs	Minimum	Median	Maximum
Observed	88	88	123	210
Predicted	88	116.9216	126.2796	168.1004

Limits of agreement (assume slope=1)

Diff:(Predicted-Observed)	2.5% Limit	97.5% Limit	SD
5.16898	-32.40471	42.74267	18.78685

Correlation between predicted and observed: 0.2873

Model 1: Full model from observed data
147.59801 +0.82329*AGE -37.61936*HEIGHT +0.30476*WEIGHT -4.38950*(OCCU.NEW=2) -
0.25422*(EDU.NEW=2) +0.01765*(EDU.NEW=3) -4.53746*(SEX=2)

	Coeff.	Se.	95%CI.low	95%CI.upp	P.value
Intercept	147.5980	25.1967	98.2124	196.9836	<0.0001
AGE	0.8233	0.0620	0.7018	0.9448	<0.0001
HEIGHT	-37.6194	17.2769	-71.4820	-3.7567	0.0298
WEIGHT	0.3048	0.1337	0.0428	0.5667	0.0229
factor(OCCU.NEW)2	-4.3895	1.5551	-7.4376	-1.3414	0.0049
factor(EDU.NEW)2	-0.2542	2.0528	-4.2776	3.7692	0.9015
factor(EDU.NEW)3	0.0176	2.4396	-4.7640	4.7993	0.9942
factor(SEX)2	-4.5375	2.3698	-9.1822	0.1073	0.0559

Model 1: predicted vs. observed **(development data)**

Summary

Method	#obs	Minimum	Median	Maximum
Observed	700	92	125	255
Predicted	700	109.7304	127.8753	165.4418

Limits of agreement (assume slope=1)

Diff:(Predicted-Observed)	2.5% Limit	97.5% Limit	SD
0.00000	-39.36859	39.36859	19.68429

Correlation between predicted and observed: 0.5187

Model 1: predicted vs. observed **(validation data)**

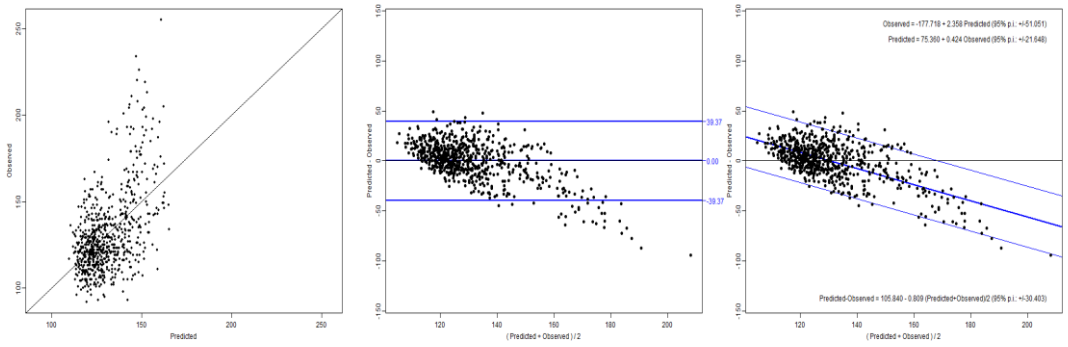
Summary

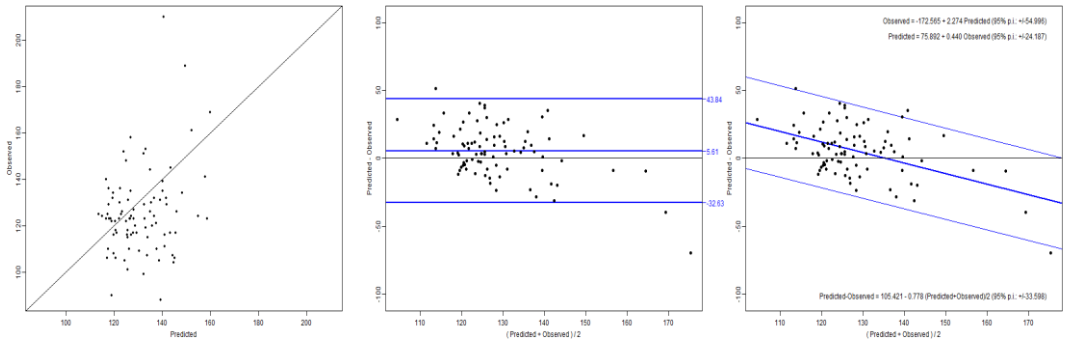
Method	#obs	Minimum	Median	Maximum
Observed	88	88	123	210
Predicted	88	113.4454	128.1342	159.9036

Limits of agreement (assume slope=1)

Diff:(Predicted-Observed)	2.5% Limit	97.5% Limit	SD
5.60501	-32.62984	43.83987	19.11743

Correlation between predicted and observed: 0.2498





Model 2: Stepwise (stepAIC) selected model from observed data
147.35379 +0.82444*AGE -37.57383*HEIGHT +0.30542*WEIGHT -4.39222*(OCCU.NEW=2) -4.50001*(SEX=2)

	Coeff.	Se.	95%CI.low	95%CI.upp	P.value
Intercept	147.3538	25.1147	98.1290	196.5786	<0.0001
AGE	0.8244	0.0544	0.7178	0.9311	<0.0001
HEIGHT	-37.5738	17.1459	-71.1798	-3.9679	0.0288
WEIGHT	0.3054	0.1332	0.0442	0.5666	0.0222
factor(OCCU.NEW)2	-4.3922	1.5115	-7.3548	-1.4296	0.0038
factor(SEX)2	-4.5000	2.2142	-8.8398	-0.1602	0.0425

Model 2: predicted vs. observed **(development data)**

Summary

Method	#obs	Minimum	Median	Maximum
Observed	700	92	125	255
Predicted	700	109.622	127.8709	165.3839

Limits of agreement(assume slope=1)

Diff:(Predicted-Observed)	2.5% Limit	97.5% Limit	SD
0.00000	-39.36932	39.36932	19.68466

Correlation between predicted and observed: 0.5187

Model 2: predicted vs. observed **(validation data)**

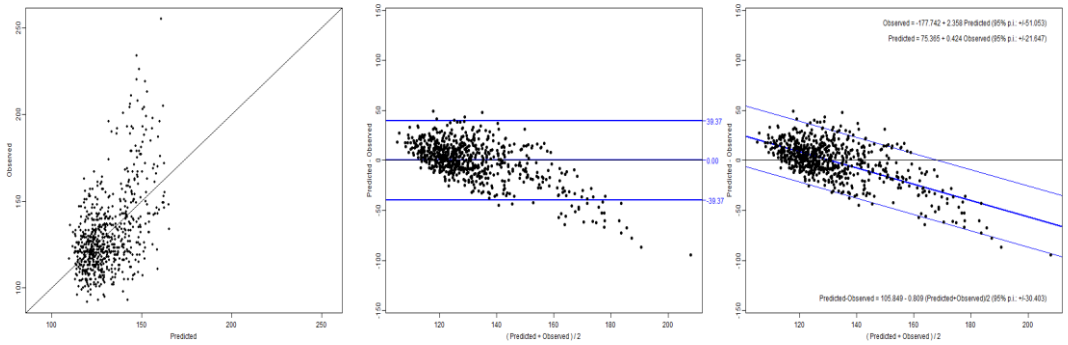
Summary

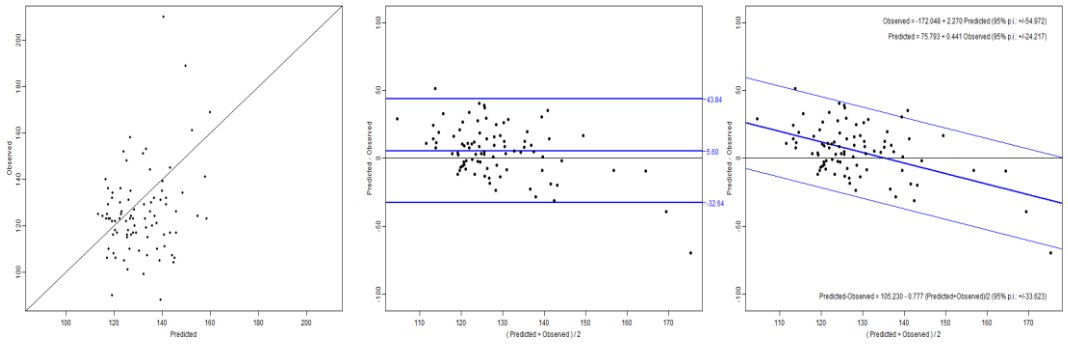
Method	#obs	Minimum	Median	Maximum
Observed	88	88	123	210
Predicted	88	113.3003	128.1625	159.8269

Limits of agreement (assume slope=1)

Diff:(Predicted-Observed)	2.5% Limit	97.5% Limit	SD
5.59979	-32.64302	43.84260	19.12140

Correlation between predicted and observed: 0.2498





例 3：练习项目 PREG 构建 PREG 的 cox 回归预测模型，输入界面如下：

输出结果：

预测模型与 ROC 分析

```
Outcome: PREG
Time: CYCLE
Collinearity VIF selection:
```

	Step 1
AGE	1
BMI	1
EDU	1
HSMK	1

Variables selected: AGE BMI EDU HSMK
Model 0: Multiple Fractional Polynomial model from observed data
Null model

Model 1: Full model from observed data
-0.00868*AGE -0.01225*BMI +0.08629*(EDU=1) +0.26896*(EDU=2) +0.13559*(HSMK=1)

	coeff.	se.	HR	Low 95%CI	High 95%CI	P value
AGE	-0.0087	0.0247	0.9914	0.9444	1.0406	0.7257
BMI	-0.0123	0.0539	0.9878	0.8888	1.0978	0.8200
factor(EDU)1	0.0863	0.1835	1.0901	0.7609	1.5619	0.6381
factor(EDU)2	0.2690	0.2567	1.3086	0.7913	2.1641	0.2947
factor(HSMK)1	0.1356	0.1608	1.1452	0.8356	1.5696	0.3992

rsq = 0.0109, maxrsq = 0.9997 , Log likelihood -814.2972

Obs	Events	Model L.R.	d.f.	P.value	Score	Score P	R2	g	gr
199	185	2.18	5	0.8241	2.2153	0.8186	0.0109	0.122	1.1298

Model 2: Stepwise selected model from observed data
NULL

预测模型 ROC 曲线分析及最佳阈值分析

Model	Time	Best.cut.X	Best.cut.sens	Best.cut.spec	Cases	N.survivor	N.censored	Cum.incidence	Surv.prob	AUC
Full	1	NA	NA	NA	0	157	42	0	0.799	NA
Full	2	-0.139196293893678	0.175	0.939	40	99	60	0.201	0.514	0.528
Full	3	0.0411122858840896	0.698	0.452	96	62	41	0.486	0.3271	0.549
Full	4	0.0252909326999191	0.621	0.523	132	44	23	0.673	0.2374	0.552

最佳阈值取敏感度+特异度最大的分界值。各分界点对应的敏感度特异保存在 ROC 输出文件（.xls）里
Inverse Probability of Censoring Weighting (IPCW) estimates of Cumulative/Dynamic time-dependent ROC curve.
注解：MFP 与 stepwise 出来的都是 null.