

Solutions for Session 7

Mark Lunt

08/11/2011

```
. do solution.do
. global basedir http://personalpages.manchester.ac.uk/staff/mark.lunt
. global datadir $basedir/stats/7_binary/data
. use $datadir/epicourse, clear
. tab hip_p sex, co
```

Key
<i>frequency</i>
<i>column percentage</i>

hip pain	sex		Total
	M	F	
no	1,969 90.16	1,976 84.77	3,945 87.38
yes	215 9.84	355 15.23	570 12.62
Total	2,184 100.00	2,331 100.00	4,515 100.00

1.1 Prevalence is 9.84% in men, 15.23% in women

. tab hip_p sex, co chi2

Key
frequency
column percentage

hip pain	sex		Total
	M	F	
no	1,969 90.16	1,976 84.77	3,945 87.38
yes	215 9.84	355 15.23	570 12.62
Total	2,184 100.00	2,331 100.00	4,515 100.00

Pearson chi2(1) = 29.6438 Pr = 0.000

1.2 The difference in prevalence between men and women is very significant

. cs hip_p sex, or

	sex		Total
	Exposed	Unexposed	
Cases	355	215	570
Noncases	1976	1969	3945
Total	2331	2184	4515
Risk	.1522952	.0984432	.1262458
	Point estimate		[95% Conf. Interval]
Risk difference	.0538519		.0346461 .0730578
Risk ratio	1.547035		1.319614 1.81365
Attr. frac. ex.	.3536024		.2422027 .4486258
Attr. frac. pop	.220226		
Odds ratio	1.645314		1.373815 1.970458 (Cornfield)

chi2(1) = 29.64 Pr>chi2 = 0.0000

1.3 Confidence interval is (1.37, 1.97)

1.4 The odds ratio and the relative risk are very similar

1.5 Yes, the confidence interval does not contain 0, which is the null hypothesis risk difference

```

. logistic hip_p sex
Logistic regression
Log likelihood = -1697.0482
Number of obs = 4515
LR chi2(1) = 29.97
Prob > chi2 = 0.0000
Pseudo R2 = 0.0088

```

hip_p	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
sex	1.645314	.1515298	5.41	0.000	1.373583 1.9708

1.6 The odds ratio is exactly the same as that produced by cs
1.7 The confidence intervals are the same to 3 decimal places (the methods used to calculate them differ, but generally give very similar results)

```

. egen agegp = cut(age), at(0 30(10)100)
. label define age 0 "<30" 30 "30-39" 40 "40-49" 50 "50-59"
. label define age 60 "60-69" 70 "70-79" 80 "80-89" 90 "90+", modify
. label values agegp age
. tab agegp hip_p, chi2

```

agegp	hip pain		Total
	no	yes	
<30	388	9	397
30-39	358	10	368
40-49	498	47	545
50-59	510	90	600
60-69	741	150	891
70-79	987	177	1,164
80-89	415	77	492
90+	48	10	58
Total	3,945	570	4,515

Pearson chi2(7) = 108.8887 Pr = 0.000

2.1 Yes: chi2 is very significant

2.5 Odds for a man aged 50-60 are 7.74 times the odds for a man aged less than 30

```
. logistic hip_p age sex
```

```
Logistic regression          Number of obs =      4515
                             LR chi2(2)          =      125.12
                             Prob > chi2         =       0.0000
Log likelihood = -1649.4709   Pseudo R2          =       0.0365
```

hip_p	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
age	1.026149	.0028811	9.19	0.000	1.020517 1.031811
sex	1.700317	.1582248	5.70	0.000	1.416839 2.040513

```
. estat gof
```

Logistic model for hip_p, goodness-of-fit test

```
number of observations =      4515
number of covariate patterns =      162
Pearson chi2(159) =      153.90
Prob > chi2 =      0.5993
```

3.1 Yes. However, this is not really appropriate, since there are so many covariate patterns. It would be better to use only 10 groups

```
. estat gof, group(10)
```

Logistic model for hip_p, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

```
number of observations =      4515
number of groups =      10
Hosmer-Lemeshow chi2(8) =      16.32
Prob > chi2 =      0.0381
```

3.1 In this case, there is evidence that the predicted and observed values differ more than can be explained by random variation

```
. lroc
```

Logistic model for hip_p

```
number of observations =      4515
area under ROC curve =      0.6368
```

```
. graph export graph1.eps replace
(file graph1.eps written in EPS format)
```

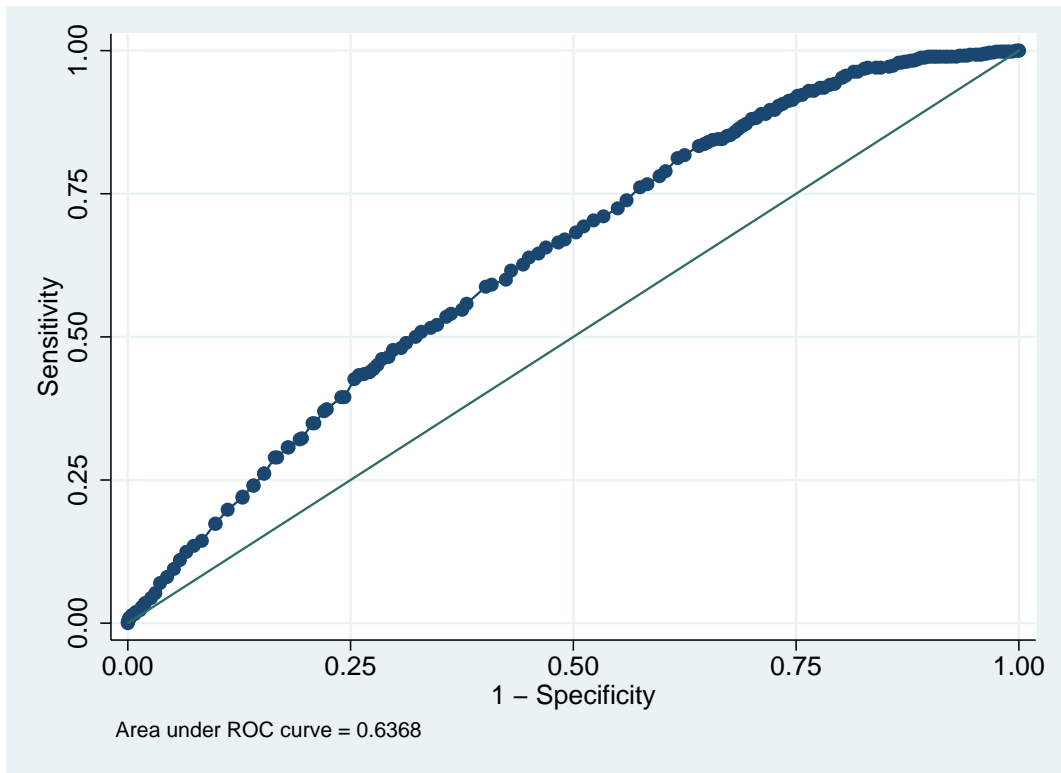


Figure 1: . lroc

```
. xi: logistic hip_p i.agegp sex
i.agegp      _Iagegp_0-90      (naturally coded; _Iagegp_0 omitted)
Logistic regression
Number of obs = 4515
LR chi2(8) = 173.58
Prob > chi2 = 0.0000
Pseudo R2 = 0.0507
Log likelihood = -1625.2427
```

hip_p	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
_Iagegp_30	1.196695	.5572489	0.39	0.700	.4804145 2.980925
_Iagegp_40	4.073759	1.509333	3.79	0.000	1.970716 8.421057
_Iagegp_50	7.740022	2.759065	5.74	0.000	3.848724 15.56567
_Iagegp_60	9.061111	3.16519	6.31	0.000	4.569233 17.96882
_Iagegp_70	7.996188	2.777287	5.99	0.000	4.047979 15.7953
_Iagegp_80	8.16381	2.937001	5.84	0.000	4.033349 16.52418
_Iagegp_90	8.463806	4.10942	4.40	0.000	3.268007 21.9204
sex	1.725543	.1611575	5.84	0.000	1.436904 2.072164

```
. estat gof
Logistic model for hip_p, goodness-of-fit test
      number of observations =      4515
      number of covariate patterns =      16
      Pearson chi2(7) =      7.94
      Prob > chi2 =      0.3381
```

```
. estat gof, group(10)
Logistic model for hip_p, goodness-of-fit test
      (Table collapsed on quantiles of estimated probabilities)
      number of observations =      4515
      number of groups =      10
      Hosmer-Lemeshow chi2(8) =      4.95
      Prob > chi2 =      0.7633
```

3.3 Yes, this model is adequate

```
. lroc
Logistic model for hip_p
      number of observations =      4515
      area under ROC curve =      0.6518
```

```
. graph export graph2.eps replace
      (file graph2.eps written in EPS format)
```

```
. gen age2= age*age
```

```
. logistic hip_p age age2 sex
```

```
Logistic regression
      Number of obs =      4515
      LR chi2(3) =      166.01
      Prob > chi2 =      0.0000
      Pseudo R2 =      0.0485
Log likelihood = -1629.0242
```

hip_p	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
age	1.164533	.0261887	6.77	0.000	1.114319	1.21701
age2	.9989405	.0001825	-5.80	0.000	.9985828	.9992983
sex	1.73314	.1616397	5.90	0.000	1.443601	2.080751

```
. estat gof, group(10)
Logistic model for hip_p, goodness-of-fit test
      (Table collapsed on quantiles of estimated probabilities)
      number of observations =      4515
      number of groups =      10
      Hosmer-Lemeshow chi2(8) =      5.12
      Prob > chi2 =      0.7442
```

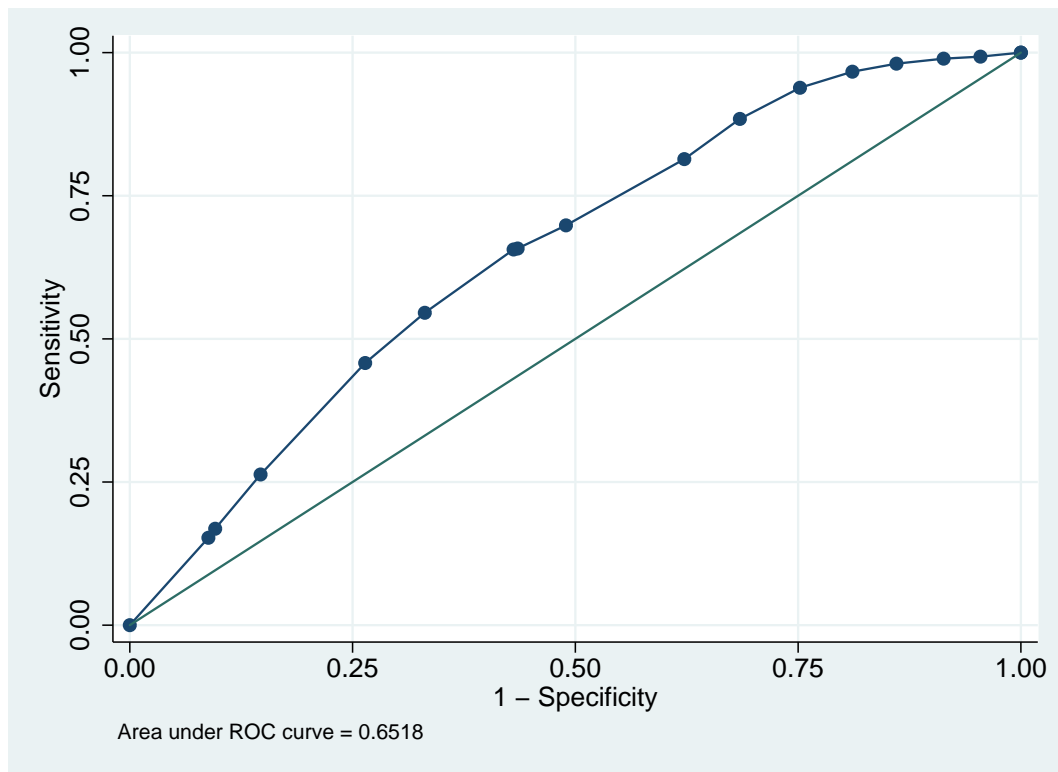


Figure 2: . lroc

3.5 Yes, the coefficient for age2 is highly significant, and there is no longer evidence of lack of fit.

```
. lroc
Logistic model for hip_p
number of observations = 4515
area under ROC curve = 0.6469

. graph export graph3.eps replace
(file graph3.eps written in EPS format)
```

3.6 The area under the curve with this model is similar to that use age as a categorical predictor.

```
. predict p
(option p assumed; Pr(hip_p))
```

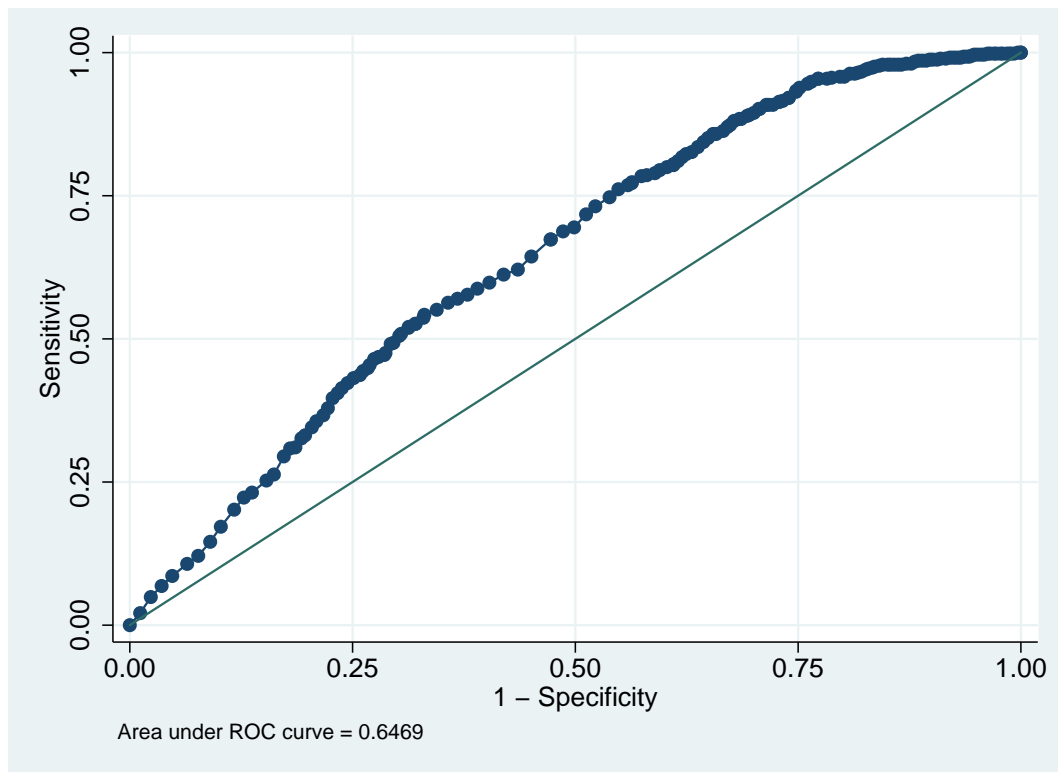


Figure 3: . lroc

```
. predict db, dbeta
. scatter db p

. graph export graph4.eps replace
(file graph4.eps written in EPS format)
```

*4.1 No, there are no points that are obvious outliers
However, there are 4 points that may be worth checking*

```
. predict d, ddeviance
. scatter d p

. graph export graph5.eps replace
(file graph5.eps written in EPS format)
```

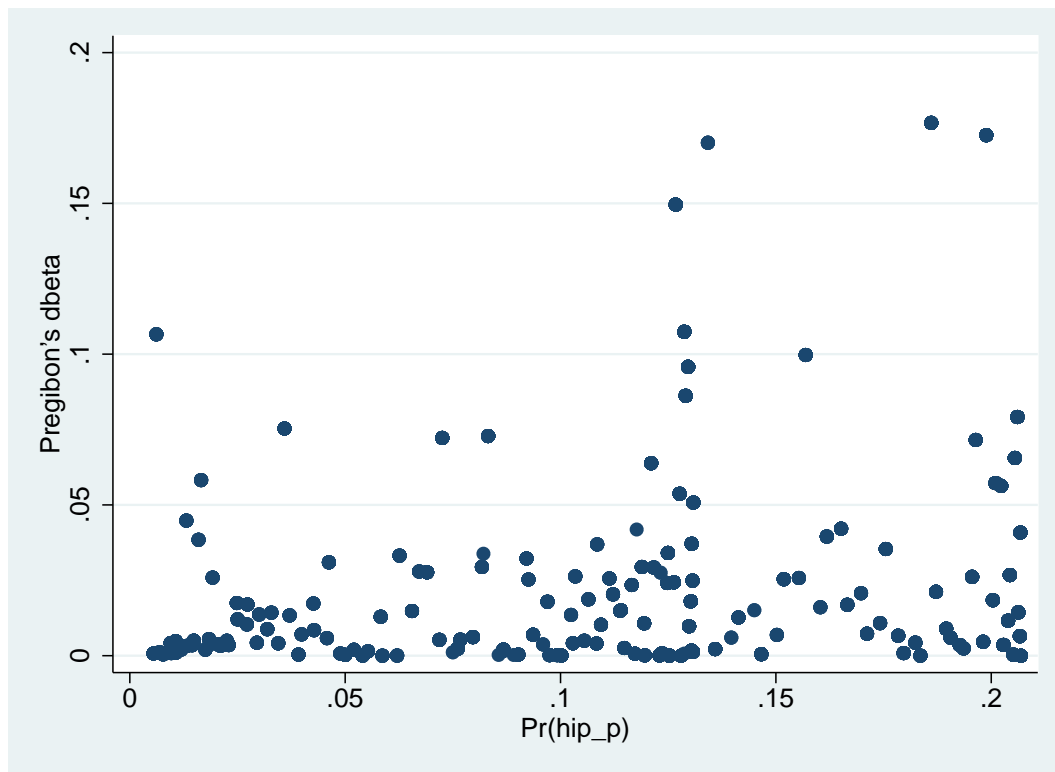


Figure 4: . scatter db p

4.2 Again, there is no evidence of any outliers

```
. scatter p age
```

```
. graph export graph6.eps replace
(file graph6.eps written in EPS format)
```

4.3 the two lines are the prevalences in men and women

```
. graph twoway scatter p age || lowess hip_p age if sex == 1 || lowess hip_p ag
e if sex == 0
```

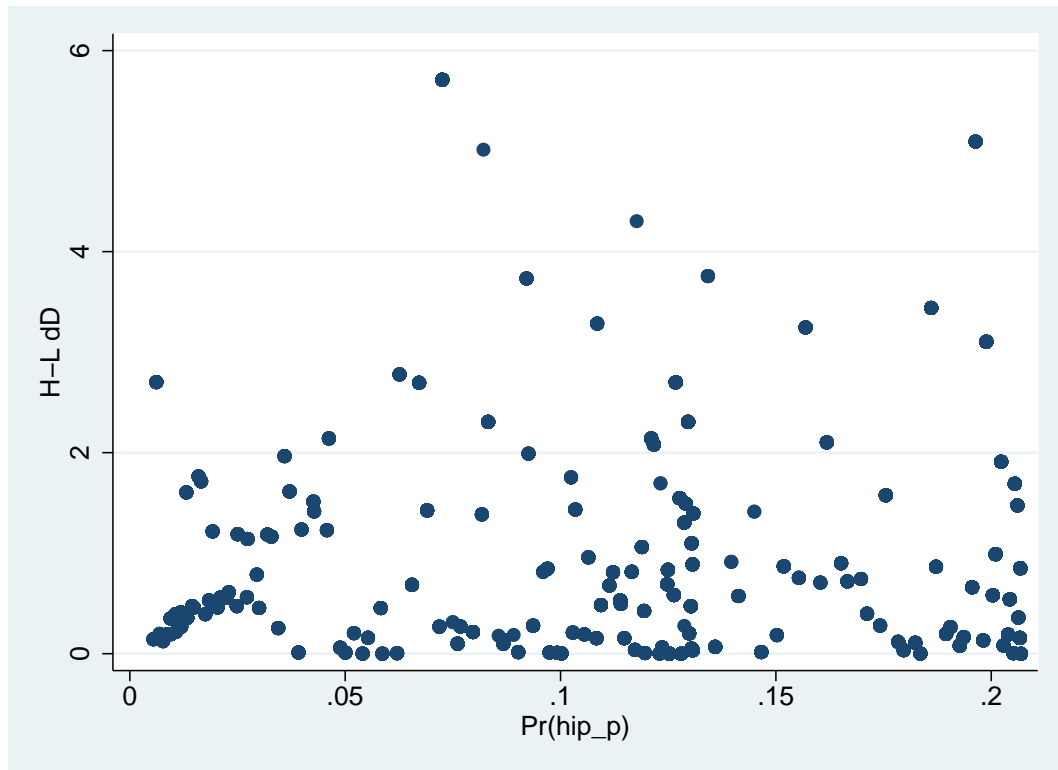


Figure 5: . scatter d p

```
. graph export graph7.eps replace
(file graph7.eps written in EPS format)
```

*4.4 the fit is good for men, but fits poorly to women over 80
The quadratic model is reasonable for men, not women*

```
. use $datadir/chd, clear
. sort agegrp
. by agegrp: egen agemean = mean(age)
. by agegrp: egen chdprop = mean(chd)
. label var agemean "Mean age"
. label var chdprop "Proportion of subjects with CHD"
. scatter chdprop agemean
```

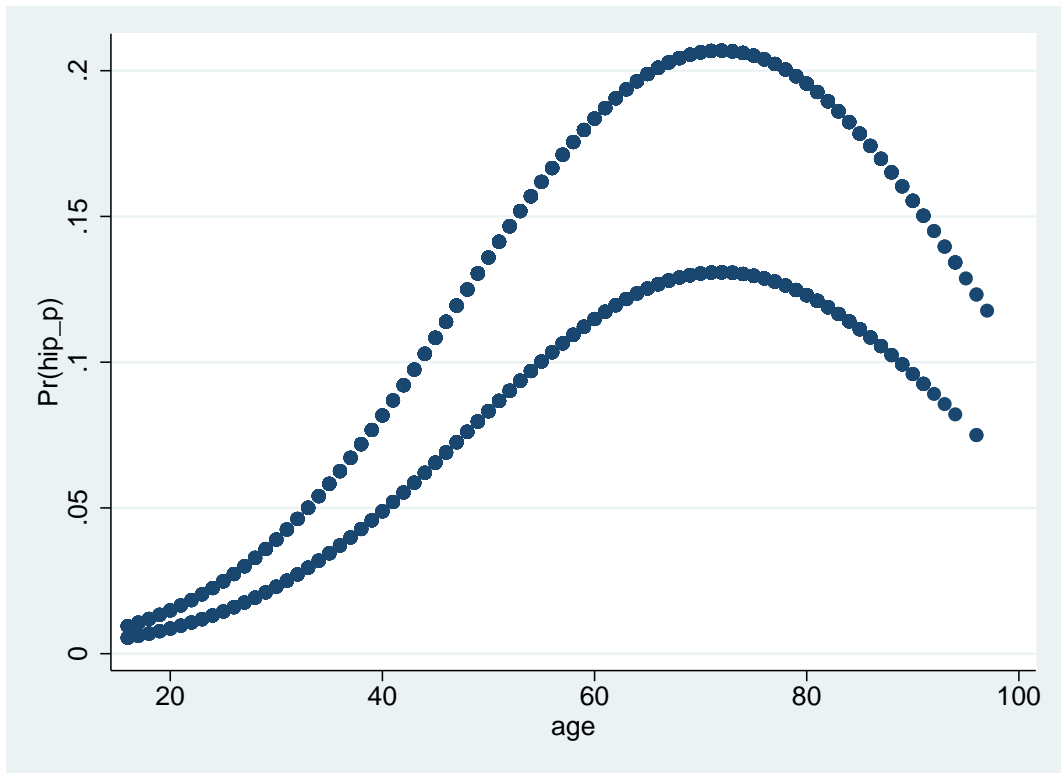


Figure 6: . scatter p age

```
. graph export graph8.eps replace
(file graph8.eps written in EPS format)
```

```
. logistic chd age
```

```
Logistic regression
Log likelihood = -53.676546
Number of obs = 100
LR chi2(1) = 29.31
Prob > chi2 = 0.0000
Pseudo R2 = 0.2145
```

	chd	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
	age	1.117307	.0268822	4.61	0.000	1.065842 1.171257

5.1 Odds ratio is about 1.12 per year

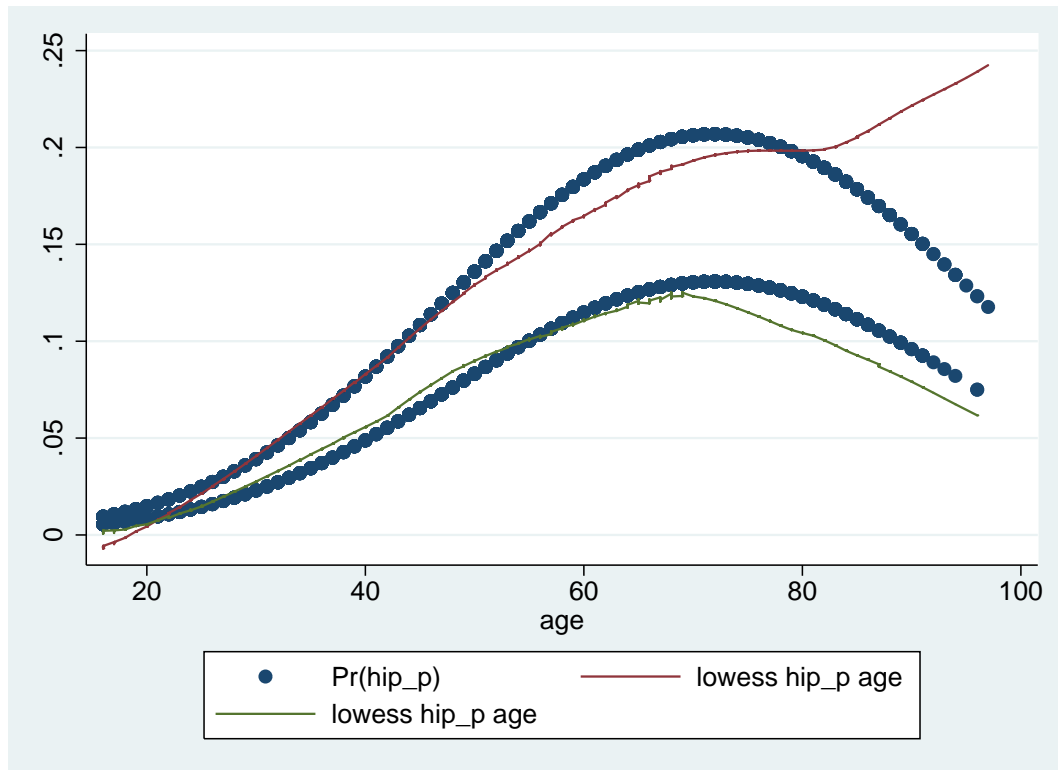


Figure 7: `. graph twoway scatter p age — lowess hip'p age if sex == 1 — lowess hip'p ag`

```
. predict p
(option p assumed; Pr(chd))

. predict db, dbeta

. predict d, ddeviance

. scatter db p

. graph export graph9.eps replace
(file graph9.eps written in EPS format)
```

5.3 Yes, there is one influential point, with db ~ 0.25

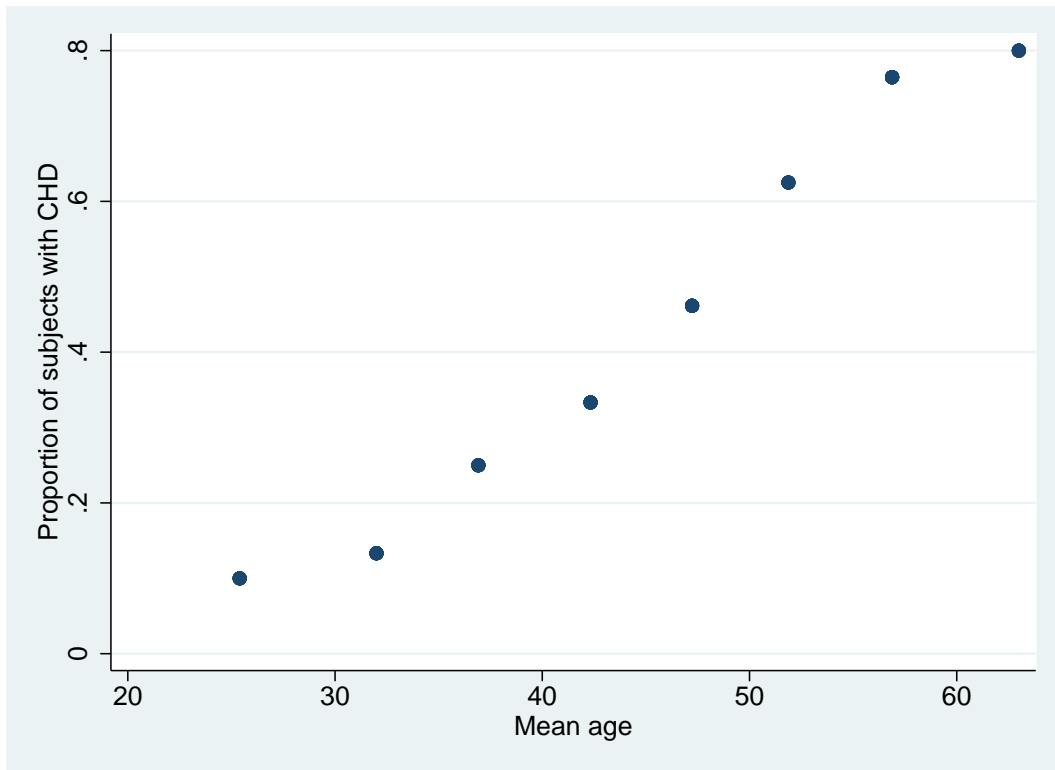


Figure 8: . scatter chdprop agemean

. summ db, detail

Pregibon's dbeta

Percentiles		Smallest		
1%	.0004816	.0004816	Obs	100
5%	.0006804	.0004816	Sum of Wgt.	100
10%	.0009566	.0004816	Mean	.0247488
25%	.0052691	.000628	Std. Dev.	.0409846
50%	.0138935		Variance	.0016797
		Largest	Skewness	3.889923
75%	.0250462	.1390621	Kurtosis	19.78814
90%	.0471688	.1390621		
95%	.0975644	.2487164		
99%	.2487164	.2487164		

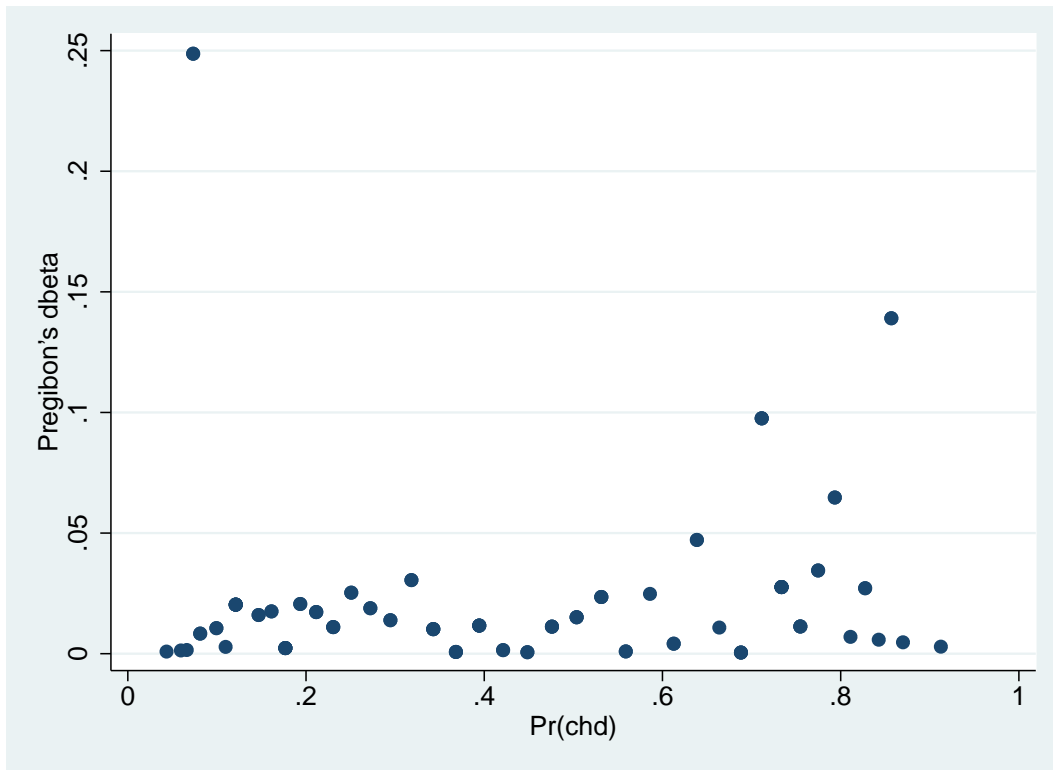


Figure 9: . scatter db p

```
. logistic chd age if db < 0.2
Logistic regression               Number of obs   =       98
                                LR chi2(1)      =       32.12
                                Prob > chi2     =       0.0000
Log likelihood = -50.863658       Pseudo R2      =       0.2400
```

	chd	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
	age	1.130329	.0293066	4.73	0.000	1.074324 1.189254

5.4 The effect on the odds ratio is small: a very slight increase to 1.13
end of do-file