

Session 6

Binary and Binomial Data

	<i>page</i>
Binary Data	6-2
Binomial	6-2
Fitting models to binary data in STATA	6-4
Parameter interpretation – logistic regression	6-12
Two-way classification of a binary response	6-16
Fitting models to binomial data in STATA	6-18
Dealing with factors in STATA	6-18
Look at parameter estimates	6-22
Plotting	6-24

Session 7: Binary and binomial data

Binary data

For each observation i , the response Y_i can take only two values coded 0 and 1.

yes/no

success/failure

presence/absence

unemployed/employed

Assume: p_i is the success probability for observation i .

y_i has a Bernoulli distribution - a special case of the Binomial distribution

Binomial

Each observation i is a count of r_i successes out of n_i trials.

Assume: p_i is the success probability for observation i .

r_i has a Binomial distribution $r_i \sim B(p_i, n_i)$

Binomial with $n_i = 1$ is Bernoulli.

Data is of the form:

r_i successes out of n_i trials

r_i is assumed to have a Binomial distribution

$$r_i \sim B(n_i, p_i)$$

1. We want to model the probability of success p_i as a function of explanatory variables.
2. We want to specify the correct distribution to carry out ML estimation, as variance of $r_i = n_i p_i (1 - p_i)$ is not constant.

Can model p_i as a linear function of explanatory variables

$$p_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots$$

Possible to get fitted values for p_i outside the range $[0,1]$.

Solution is to transform the success probability.

If $H(\theta)$ as an increasing function of θ

$$H(-\infty) = 0 \quad H(\infty) = 1$$

Then $H(\cdot)$ defines transformations from $(-\infty, \infty)$ to $(0,1)$.

Example: $H(\cdot)$ can be any cumulative distribution function defined on $(-\infty, \infty)$.

e.g. Normal $H(\cdot) = \Phi(\cdot)$

Define **LINEAR PREDICTOR** η_i to be $\underline{\beta}' \underline{X}_i$

Then

$$p_i = H(\eta_i)$$

$$E(r_i) = n_i H(\eta_i)$$

Inverse of $H(\cdot)$ is called the **LINK FUNCTION** $g(\cdot)$

$$g(\cdot) = H^{-1}(\cdot)$$

$$g(p_i) = \eta_i$$

Example: LOGIT LINK

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \underline{\beta}' \underline{X}_i$$

$$= \eta_i$$

$$p_i = H(\eta_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

$H(\cdot)$ is c.d.f for logistic distribution.

Example PROBIT LINK

$$g(p_i) = \Phi^{-1}(p_i) = \underline{\beta}' \underline{X}_i$$

$$p_i = H(\eta_i) = \Phi(\eta_i)$$

$H(\cdot)$ is c.d.f for Normal distribution.

Fitting models to binary data in STATA

Can use `glm` command or wide range of specialist commands:

logit link- binary data

`logistic` 'logistic regression'
`logit` 'maximum likelihood logit regression'

logit link – binomial data

`glogit` 'logit on grouped data'
`blogit` 'weighted least squares estimates for grouped data'

probit link- binary data

`probit` 'maximum likelihood logit regression'

probit link – binomial data

`gprobit` 'probit on grouped data'
`bprobit` 'weighted least squares estimates for grouped data'

For logit link, binary data `logit` and `logistic` command are similar

`logit` response-variable explanatory vars

Statistics>Binary Outcomes>logistic regression

Example VASO-CONSTRICTION data

Finney(1947, Biometrika)

Response is vasoconstriction in the skin of the fingertips.
RESP

Explanatory variables are two continuous variables:

VOL – volume of air inhaled

RATE – rate of air inhaled.

39 observations – only 3 subjects, but ignore this for now.

D. J. FINNEY

321

2. THE DATA

Research in human physiology has demonstrated that, under carefully controlled experimental conditions, a transient reflex vaso-constriction in the skin of the digits may follow a single deep breath (Bolton, Carmichael & Stürup, 1936). Gilliatt (1947) has found that the response depends in part on the volume of air taken in by the subject. Plethysmographic measurement of the volume changes in a finger was used to indicate the occurrence of a response, but assessment of the degree of vaso-constriction, in order to relate this to the inspiratory stimulus, was not practicable. Thus the records obtained for each test show only

the volume of air inspired, the average rate of inspiration, and whether or not vaso-constriction was produced. The above brief outline is sufficient for appreciation of the statistical problem, but a full account of the experimental procedure may be found in Gilliatt's paper; the results discussed here are presented in his Fig. 5.

The data, which Mr Gilliatt has kindly made available to the writer, were obtained from thirty-nine tests, in twenty of which vaso-constriction occurred. Tests were made on three different subjects, nine on D.W., eight on V.P.W., and twenty-two on S.J.S.; the results of the tests, with the subjects in this order, are shown in Table 1. In Fig. 1 are shown the thirty-nine combinations of volume in litres (V) and rate of inspiration in litres per second

Overleaf we see Finney's plot.

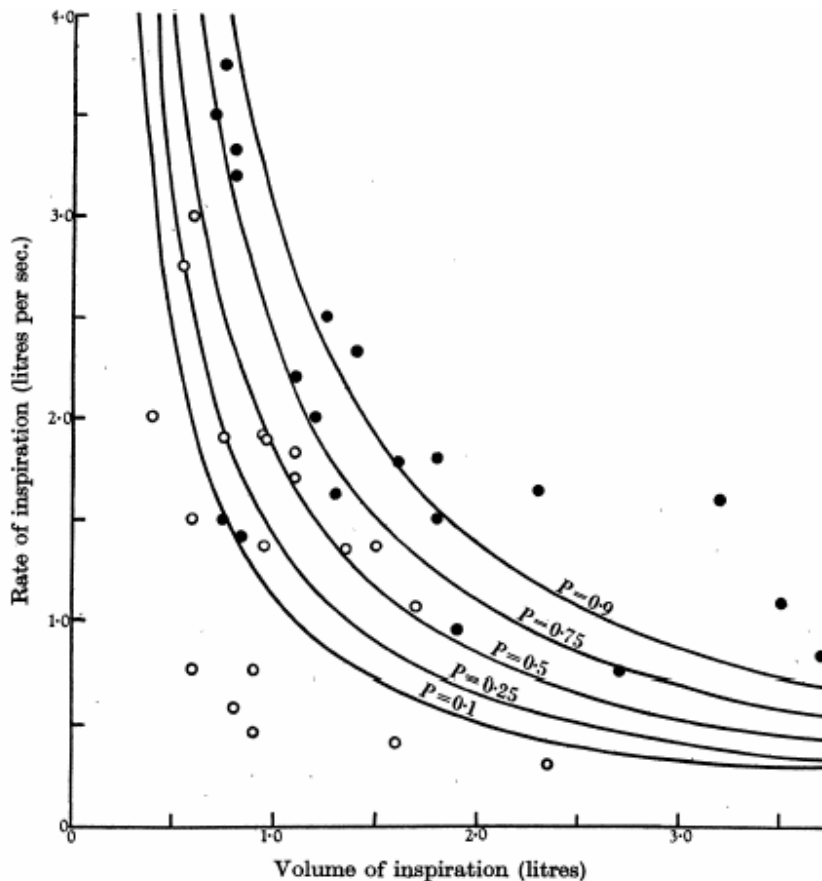


Fig. 1. Contours of dose-response surface for 0.1, 0.25, 0.5, 0.75 and 0.9 frequency of response, estimated from three-parameter equation. ○ no vaso-constriction; ● vaso-constriction.

For one point, the data published in the paper does not agree with the plot. This is point 32. The value of RATE given in the paper is 0.03, but in the plot, it appears closer to 0.3. Finney did his calculations by hand and did not use a computer, but it appears that 0.3 is the correct value. We have therefore modified the data.

The plot shows a strong relationship between both RATE and VOL, with the probability of vasoconstriction increasing as either or both increase.

We fit a logistic regression in STATA.

```
logit RESP RATE VOL
```

Following logit command can use `predict` : various vectors produced by fitting.

. logit RESP RATE VOL

Iteration 0: log likelihood = -27.019918
Iteration 1: log likelihood = -17.183044
Iteration 2: log likelihood = -15.570635
Iteration 3: log likelihood = -15.246015
Iteration 4: log likelihood = -15.228512
Iteration 5: log likelihood = -15.228447

Logit estimates

Number of obs = 39
LR chi2(2) = 23.58
Prob > chi2 = 0.0000
Pseudo R2 = 0.4364

Log likelihood = -15.228447

RESP	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
RATE	2.592717	.9058165	2.86	0.004	.8173495	4.368085
VOL	3.66041	1.333405	2.75	0.006	1.046985	6.273835
_cons	-9.186611	3.10418	-2.96	0.003	-15.27069	-3.102531

RESP	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
RATE	13.36604	12.10718	2.86	0.004	2.26449	78.89242
VOL	38.87728	51.83914	2.75	0.006	2.849049	530.5079

Differences in scaled deviances for two models, with one a submodel of the other, have a χ^2 distribution with K df, if the K parameters omitted are zero.

Omit RATE from main effects model

Or

Omit VOL from main effects model

Test against $\chi^2_1 \xrightarrow{-}$ both RATE and VOL are important.
Compare with critical value of chi-squared distribution.

$$\chi^2_{0.05,1} = 3.84$$

parameter	estimate	s.e.	z
1	-9.186	3.104	-2.96
RATE	2.593	0.906	2.86
VOL	3.660	1.333	2.75

Approximate test to indicate likely terms to be excluded is to look at the 'z-values' (estimate/s.e.). If (estimate/s.e.) is small (less than 2), then most likely good candidate for removal.

Or look at $P > |z|$ - if less than 0.05, then not candidate for removal.

No candidates identified here – model can not be simplified.

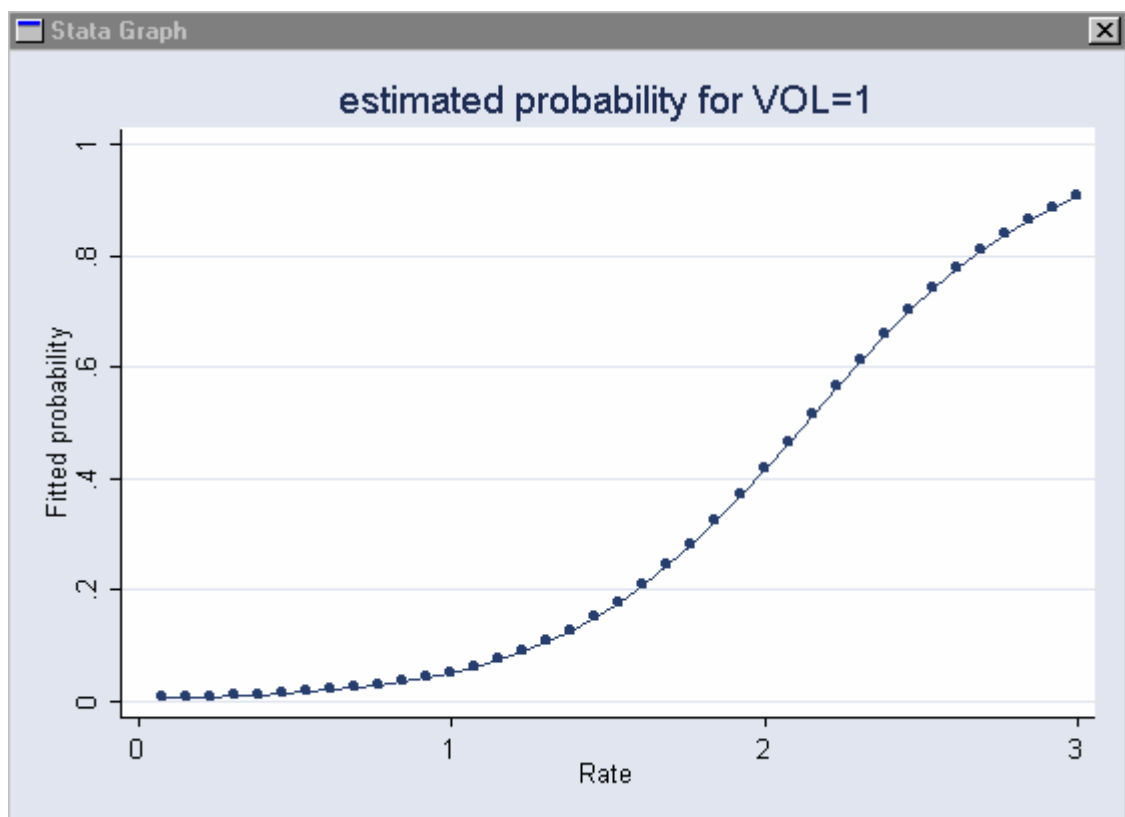
Change in scaled deviance should then be calculated.

For fixed VOL, what is relationship between probability of vaso-restriction and RATE?

For VOL=1, calculate fitted probabilities over range of values of RATE.

```
gen r=sum(1)/13
gen lp=-9.187 + 3.660*1 + 2.593*r
gen elp=exp(lp)
gen fp=elp/(1+elp)
```

```
twoway (connected fp r),
       ytitle(Fitted probability)
       xtitle(Rate)
       title(estimated probability for VOL=1)
```



Parameter interpretation – logistic regression

$$\log\left(\frac{p}{1-p}\right) = -9.187 + 3.660 * VOL + 2.593 * RATE$$

- For fixed RATE, the effect of a unit increase in VOL is to increase the log-odds by 3.660.
- For fixed RATE, the effect of a unit increase in VOL is to multiply the odds of vaso-constriction by $\exp(3.660) = 38.88$

95% confidence intervals (C.I.) for odds are often calculated in medical reports.

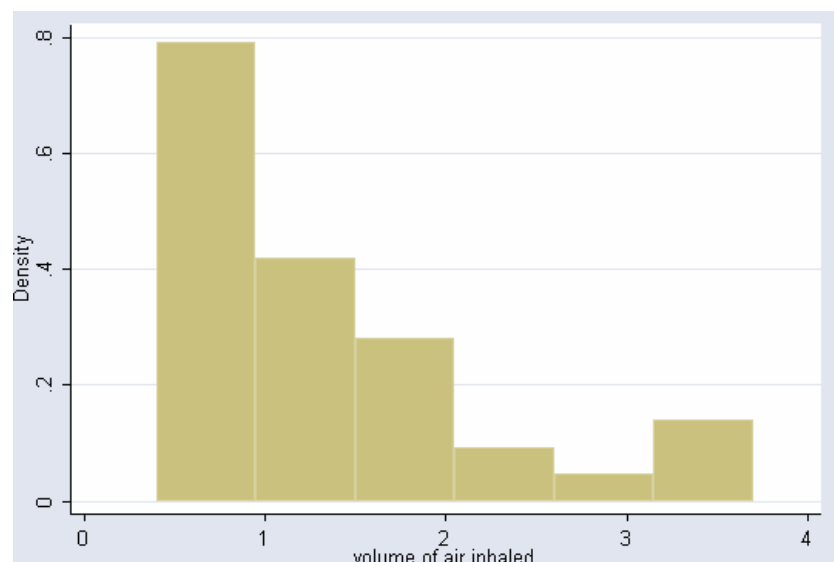
If C.I. for odds contains 1.0, then no evidence that covariate is important.

C.I for parameter estimate for VOL is

$$(3.660 - 1.96 * 1.333, 3.660 + 1.96 * 1.333)$$
$$(1.047, 6.274)$$

C.I for VOL odds is

$$(\exp(3.660 - 1.96 * 1.333), \exp(3.660 + 1.96 * 1.333))$$
$$(\exp(1.047), \exp(6.274))$$
$$(2.85, 530.51)$$



extracting fitted values and residuals

```
predict fv          store fitted probabilities in fv  
predict res,r      store pearson residuals in res
```

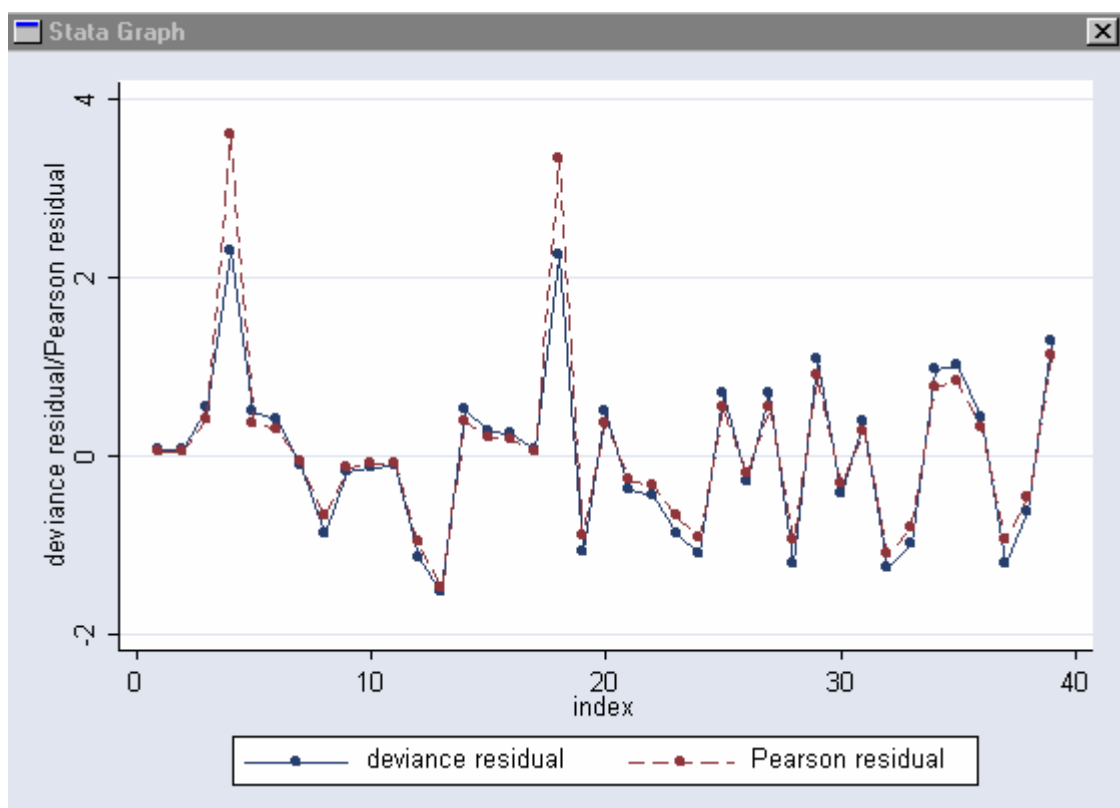
Pearson residuals defined by

$$(y_i - \hat{p}_i) / [\hat{p}_i(1 - \hat{p}_i)]^{1/2}$$

```
predict dev,de  
deviance residuals -signed contribution to scaled deviance
```

two large residuals -4th and 18th observations.

```
Two way overlay graph  
twoway (connected dev index)  
      (connected res index, clpat(dash))
```



Try other models?

1. increase complexity of model

Fit interaction between RATE and VOL

```
gen RV=RATE*VOL
logit RESP RATE VOL
est store A
logit RESP RATE VOL RV
lrtest A
```

log-likelihood = -13.36 scaled deviance = 26.71
change from main effects model = 30.46-26.71
= 3.74 on 1df

Borderline significant (p=0.053)

2. try transformation of explanatory variables

```
gen LVOL=log(VOL)
gen LRATE=log(RATE)
logistic RESP LVOL LRATE
```

log-likelihood=-14.63 scaled deviance= 29.26
slight but no great improvement. We prefer simpler
interpretation of untransformed model.

3. Try different link function

PROBIT:

```
probit RESP VOL RATE
```

Probit estimates

```
Number of obs   =          39  
LR chi2(2)      =          23.40  
Prob > chi2     =          0.0000  
Pseudo R2      =          0.4331
```

Log likelihood = -15.317606

RESP	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
VOL	2.022317	.6690106	3.02	0.003	.7110804	3.333554
RATE	1.455868	.4599026	3.17	0.002	.5544753	2.35726
_cons	-5.060134	1.496411	-3.38	0.001	-7.993045	-2.127223

scaled deviance=30.63

Fit slightly poorer than logit link.
Interpretation is also harder.

Two-way Classification of a binary response

Study of coronary heart disease

1329 males classified by

- serum cholesterol
- systolic blood pressure

diagnosed with coronary heart disease. (yes/no)

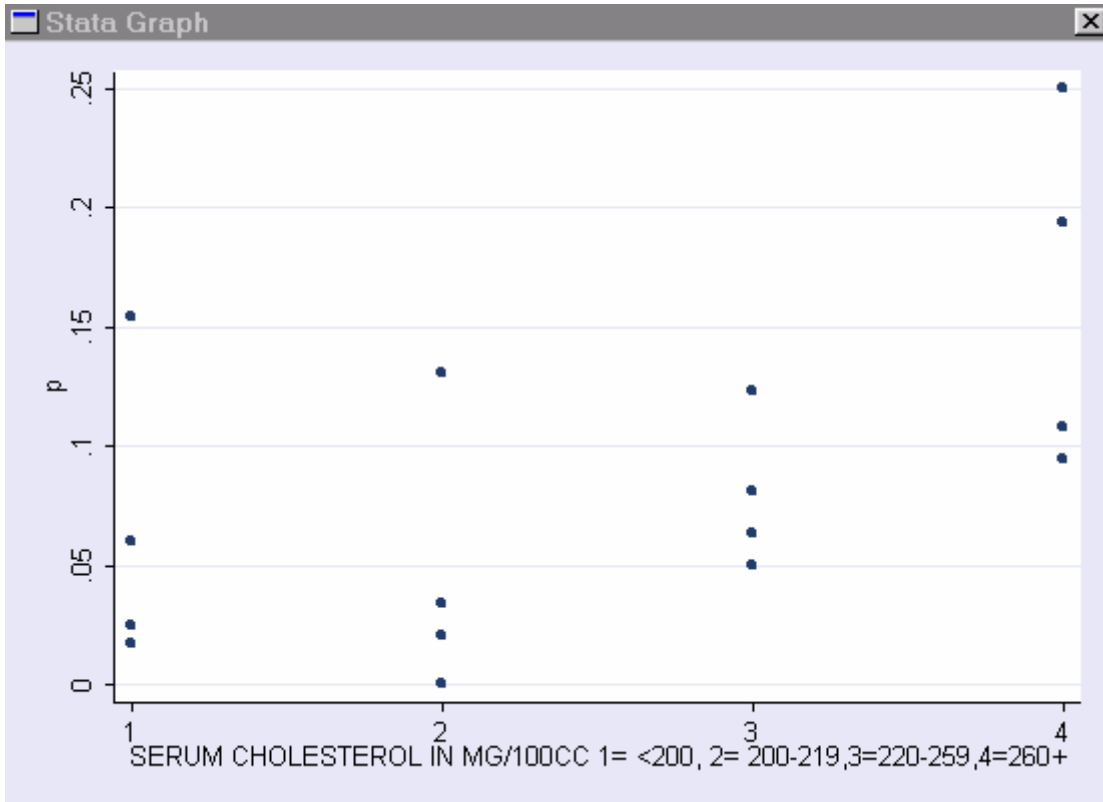
Date from Ku and Kullback American Statist. 1974

		Blood pressure			
		<127	127-146	147-166	>=167
serum cholest	<200	2/119	3/124	3/50	4/26
	200-219	3/88	2/100	0/43	3/23
	220-259	8/127	11/220	6/74	6/49
	>259	7/74	12/111	11/57	11/44

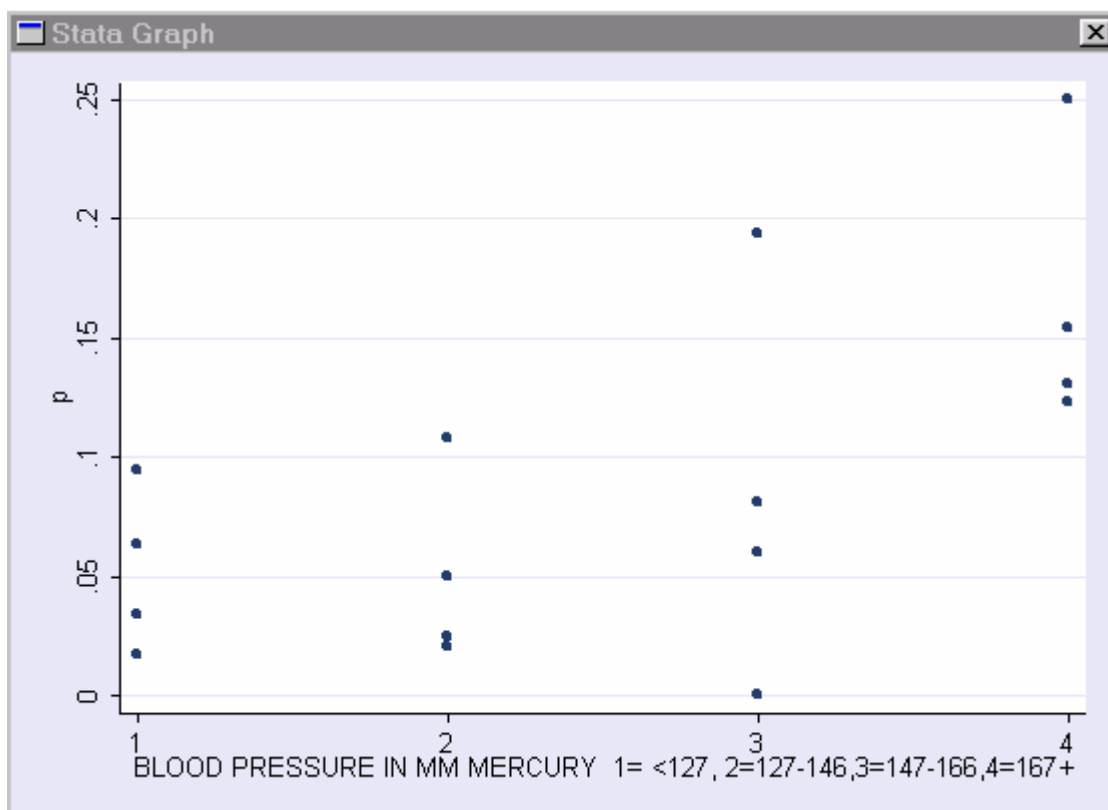
r	number suffering from heart disease	
n	total	
chol	serum cholesterol	} treat as unordered factors
bp	blood pressure	

1. Plot proportions suffering from heart disease against cross-classifying factors:

```
gen p=r/n  
twoway scatter p chol  
twoway scatter p bp
```



generally increasing P with levels of each factor.



Recall...

Fitting models to Binomial data in STATA

Can use `glm` command or use specialist commands:

logit link – binomial data

`blogit...` 'maximum likelihood logit on grouped data'

`glogit` weighted least squares estimates for
 grouped data

probit link – binomial data

`bprobit` 'maximum likelihood probit on grouped
 data'

`gprobit` weighted least squares estimates for
 grouped data

We use `blogit` or `bprobit`

Dealing with factors in STATA

We need to get STATA to form dummy variables out of the factors BP and CHOL

We use the `xi:` prefix command to all fitting commands and use the term `i.factor` to include factors in model. This can not be done through the graphical front end.

```
xi: blogit r n i.bp i.chol
```

```
xi: blogit r n i.bp i.chol
i.bp          Ibp_1-4      (naturally coded; Ibp_1 omitted)
i.chol        Ichol_1-4    (naturally coded; Ichol_1 omitted)
```

```
Logit Estimates                                     Number of obs =   1329
                                                    chi2(6)          =   50.65
                                                    Prob > chi2      = 0.0000
Log Likelihood = -309.09068                          Pseudo R2       = 0.0757
```

_____	_____	_____	_____	_____	_____	_____
_outcome	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+	-----	-----	-----	-----	-----	-----
Ibp_2	-.0414608	.3036517	-0.137	0.891	-.6366072	.5536855
Ibp_3	.5323561	.3323976	1.602	0.109	-.1191311	1.183843
Ibp_4	1.200422	.3268887	3.672	0.000	.5597321	1.841112
Ichol_2	-.2079774	.4664193	-0.446	0.656	-1.122142	.7061875
Ichol_3	.5622288	.3507979	1.603	0.109	-.1253224	1.24978
Ichol_4	1.344121	.3429662	3.919	0.000	.6719193	2.016322
_cons	-3.481939	.3486498	-9.987	0.000	-4.16528	-2.798598
-----	-----	-----	-----	-----	-----	-----

Scaled deviance for grouped binary data

$$= -2 \log \text{likelihood (current model)} - (-2 \log \text{likelihood (saturated model)})$$

Saturated model is the model where there is a parameter for every observation
 Model reproduces data exactly.

Saturated model provides a baseline for assessing values of likelihood. We set baseline through `est` command

a) fit saturated model (all two-way interaction model)

```
xi: blogit r n i.chol i.bp i.chol*i.bp  
warning message!!  
Note: IcXb_2_3!=0 predicts failure  
perfectly  
IcXb_2_3 dropped and 1 obs not used
```

Repeat using `asis` option

```
xi: blogit r n i.chol i.bp i.chol*i.bp,  
asis
```

Now no warning message, but stata gets df wrong!

b) store likelihood

```
est store sat
```

c) fit any other model (eg main effects)

```
xi: blogit r n i.chol i.bp i.chol.bp
```

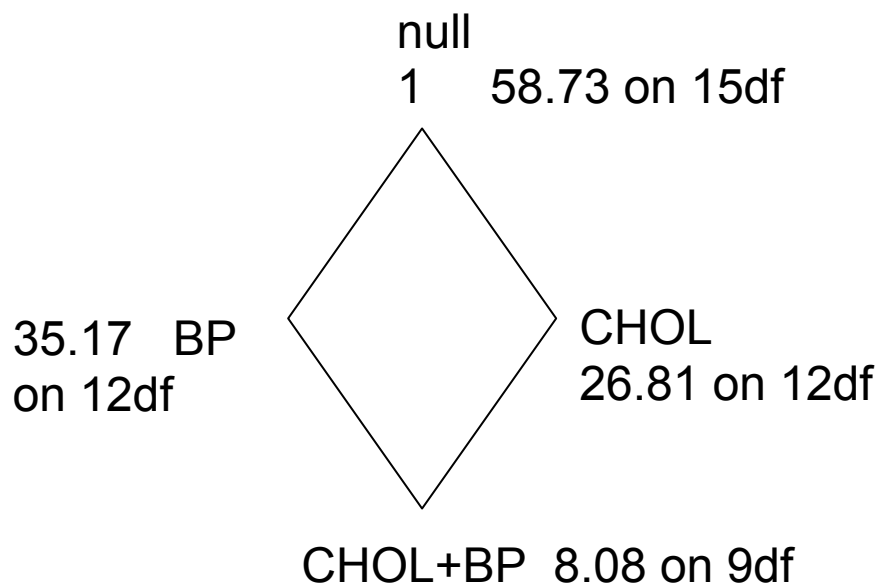
d) carry out likelihood ratio test with saturated model.

```
lrtest sat
```

```
likelihood-ratio test    LR chi2(8) = 8.08  
(Assumption: . nested in sat)  
Prob > chi2 =          0.4261
```

Scaled deviance is 8.08 on 9 degrees of freedom
(note STATA gets the degrees of freedom wrong)

4. Scaled deviances are:



Both CHOL and BP important

Main effects model provides good fit to the data.

8.08 on 9 df consistent with χ^2_9

Test valid if all N large.

Examine residuals

Harder with grouped data as STATA does not provide them.

```
predict fitp
gen res=(r-n*fitp)/sqrt(n*fitp*(1-
fitp))
list res
```

- none large

Look at parameter estimates

1	-3.482
CHOL(2)	-0.208
(3)	0.562
(4)	1.344
BP(2)	-0.042
(3)	0.532
(4)	1.200

Consistent increase with factor level for both factors

(a) try CHOL and BP as continuous scores

```
blogit r n chol bp
lrtest sat
```

Scaled deviance is 14.85 on 13 df – model still fits (p=0.25)

Scaled deviance change of 6.77 on 4 df.

```
drop res
predict fitp
gen res=(r-n*fitp)/sqrt(n*fitp*(1-fitp))
```

large residual – unit 4

(b) try combining levels 1 and 2 of CHOL and BP then fit as continuous scores

```
create new variables CH and B
gen ch=chol
gen b=bp
```

1	1
2 →	1
3	2
4	3

```

recode b 1 2 =1 2=1 3=2 4=3
recode ch 1 2 =1 2=1 3=2 4=3

blogit r n ch b
lrtest sat

```

Scaled deviance is 8.42 on 13 df
Change from main effects factor model is 0.34 on 4 df.

	estimate	s.e
CH	0.72	0.14
B	0.61	0.13

$$\beta_0 + \beta_1 CH + \beta_2 B$$

Can think of constraining the estimate of CH to be equal to that of B.

$$\beta_0 + \beta_1' CH + \beta_1' B$$

$$\beta_0 + \beta_1' (CH + B)$$

```

gen bch = b+ch
blogit r n bch

```

Scaled deviance is now 8.74 on 14 df.

	estimate	s.e.
BCH	0.66	0.12

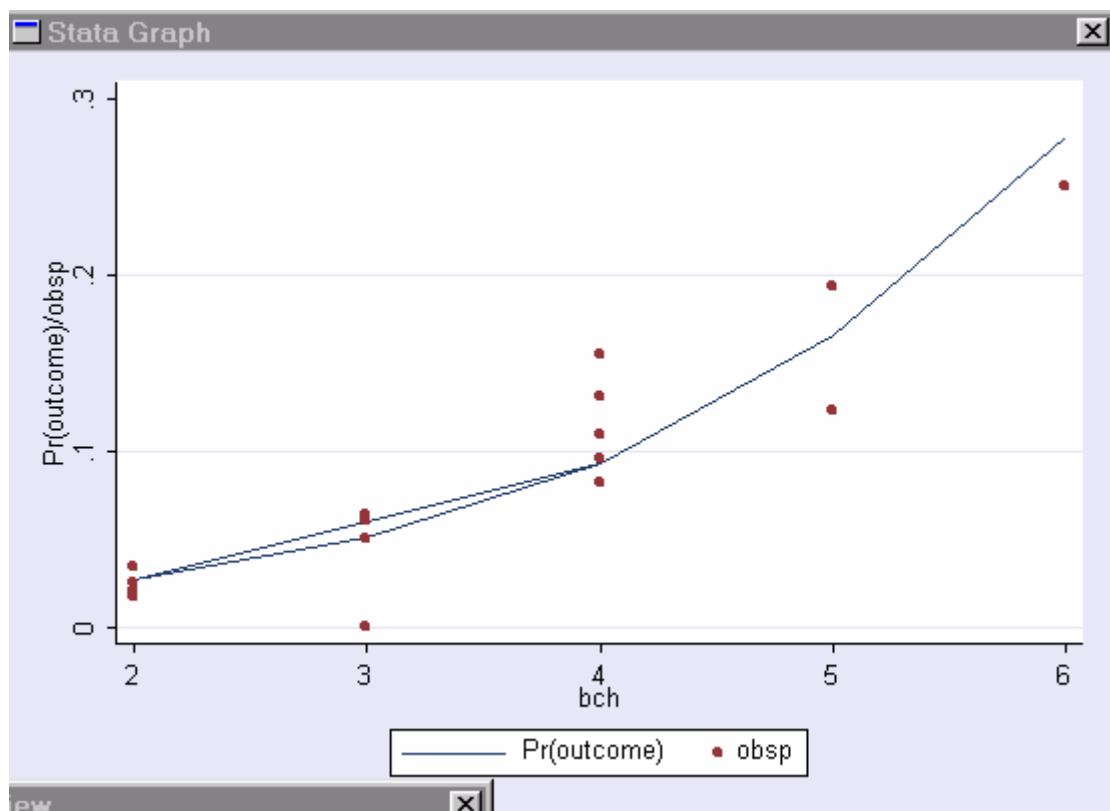
$$e^{0.66} = 1.93 \text{ nearly } 2!$$

Log-odds of coronary heart disease doubles with unit increase of BCH. BCH can be thought of as a risk score.

Plotting

Now only 5 values of BCH, rather than 16 categories.

```
drop fitp
predict fitp
twoway (line fitp bch) (scatter obsp
bch)
```



Conclusion. Excellent final model, but beware of saturated models in STATA. Take care and check the degrees of freedom.