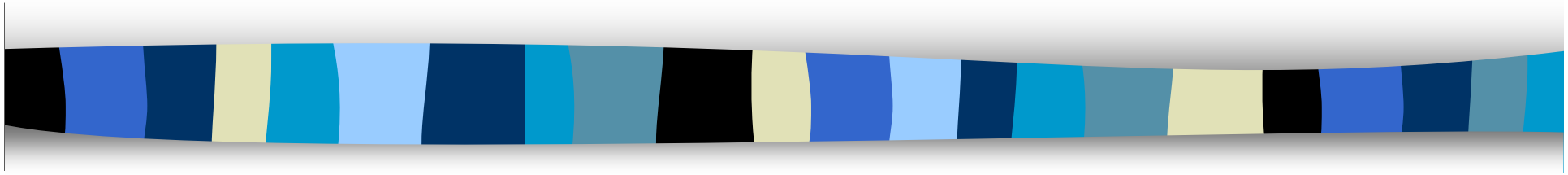
SSSII

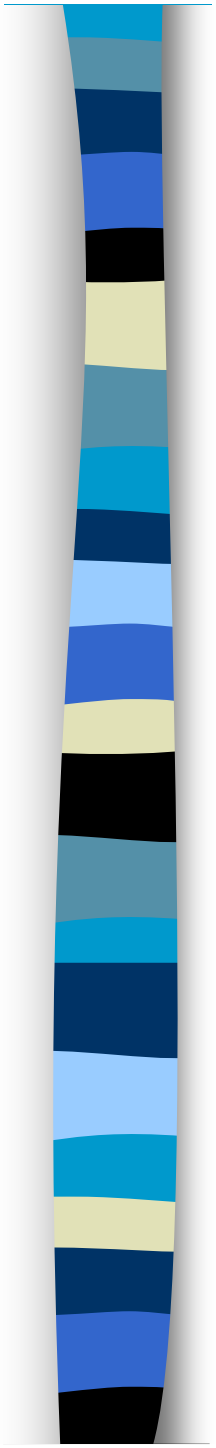
Gwilym Pryce



Lecture 8: Binary Dependent Variable Estimation

Notices:

- Register
- Lecture starts later next week:
  - 2.10pm to 3pm: Intro to Lab 5 Omitted variables;
    - » plus questions on the assignment etc.
- SSSG: email me if you want to be added to the mailing list.
- QIMF
- For a variety of statistical resources, see AQMeN
  - – http://aqmen.ac.uk/



Schedule:

- Mon 28th Feb:
  - Lecture 1pm-2.30pm
    - topic 8
    - Questions re assignment
  - Lab
    - Lab 4 cont., + catch-up
- Mon 7th March
  - Lecture 2pm-3pm
    - Intro to lab 5
    - Questions re assignment
  - Lab
    - Lab 5 (2hrs)

Plan:

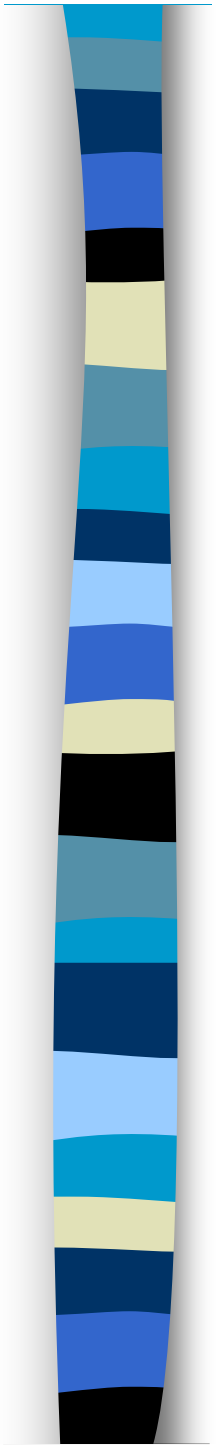
- 1. Overview of Non-Continuous Dependent Variables
- 2. Linear Probability Model
- 3. Logit
- 4. Logit Estimation & Interpretation
- 5. Multiple Logit Regression

1. Overview of Non-continuous Dependent Variables

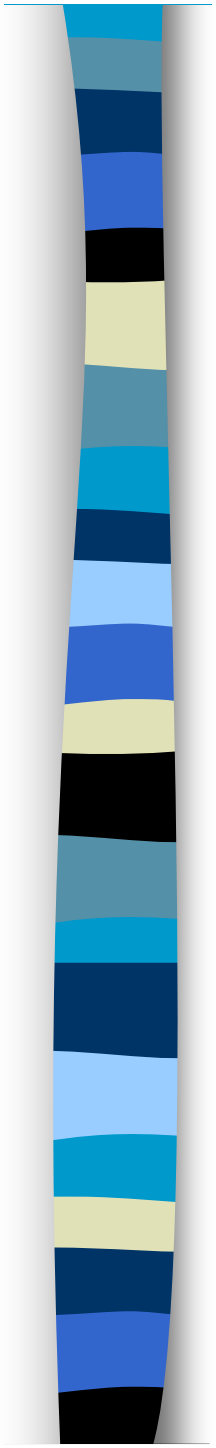
- linear regression model: most commonly used statistical tool in the social sciences
- but it assumes that the dependent variable is an uncensored, unbounded, “scale numeric” variable
  - I.e. it is continuous and has been measured for all cases in the sample
- however, in many situations of interest to social scientists, the dependent variable is not continuous or measured for all cases (taken from Long, 1997, p. 1-3):

Q/ What types of variable might the dependent variable be, other than continuous?

- Give examples



- e.g. 1 **Binary variables:** made up of two categories
  - coded 1 if event has occurred, 0 if not.
  - It has to be a decision or a category that can be explained by other variables (I.e. male/female is not something amenable to social scientific explanation -- it is not usually a *dependent* variable):
    - Did the person vote or not?
    - Did the person take out MPPI or not?
    - Does the person own their own home or not?
  - *If the Dependent variable is Binary then Estimate using:* binary logit (also called logistic regression) or probit

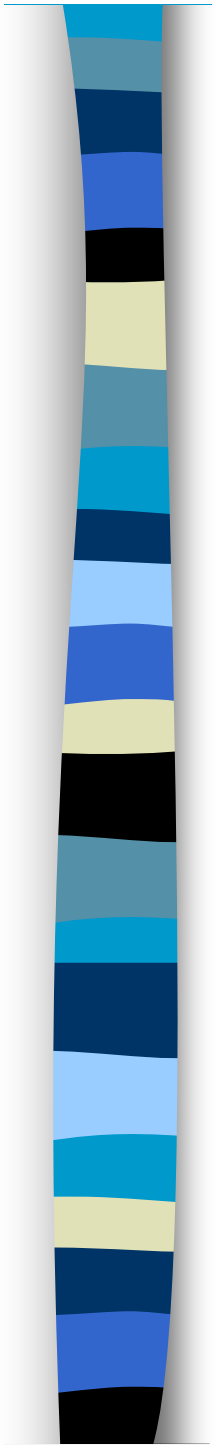


- e.g. 2 **Ordinal variables:** made up of categories that can be ranked (ordinal = “has an inherent order”)
  - e.g. coded 4 if strongly agree, 3 if agree, 2 if disagree, and 1 if strongly disagree.
  - e.g. coded 4 if often, 3 occasionally, 2 if seldom, 1 if never
  - e.g. coded 3 if radical, 2 if liberal, 1if conservative
  - e.g. coded 6 if has PhD, 5 if has Masters, 4 if has Degree, 3 if has Highers, 2 if has Standard Grades, 1 if no qualifications
  - *If the Dependent variable is Ordinal then Estimate using:* ordered logit or ordered probit

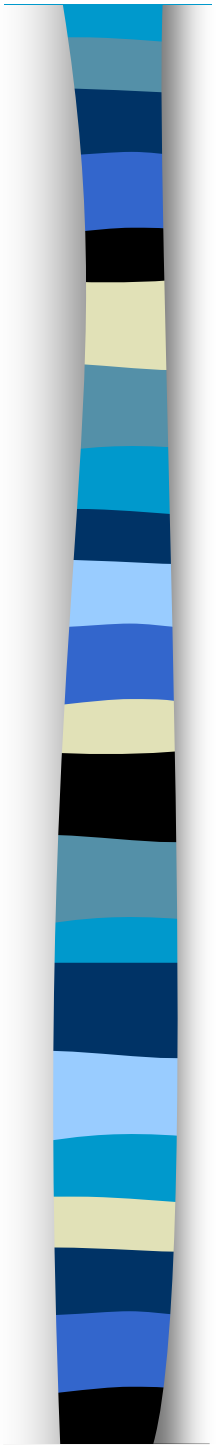
- e.g.3 **Nominal variables:** made up of multiple outcomes that cannot be ordered
  - e.g. Marital status: single, married, divorced, widowed
  - e.g. mode of transport: car, van, bus, train, bycycle
  - *If the Dependent variable is Nominal then Estimate using:* multinomial logit

- e.g. 4 **Count variables:** indicates the number of times that an event has occurred.
  - e.g. how many times has a person been married
  - e.g. how often times did a person visit the doctor last year?
  - e.g. how many strikes occurred?
  - e.g. how many articles has an academic published?
  - e.g. how many years of education has a person completed?
  - *If the Dependent variable is a Count variable Estimate using:* Poisson or negative binomial regression

- E.g 5 **Censored Variables**: occur when the value of a variable is unkown over a certain range of the variable
  - e.g. variables measuring %: censored below at zero and above at 100.
  - e.g. hourly wage rates: censored below by minimum wage rate.
  - *If the Dependent variable is Censored, Estimate using:* Tobit

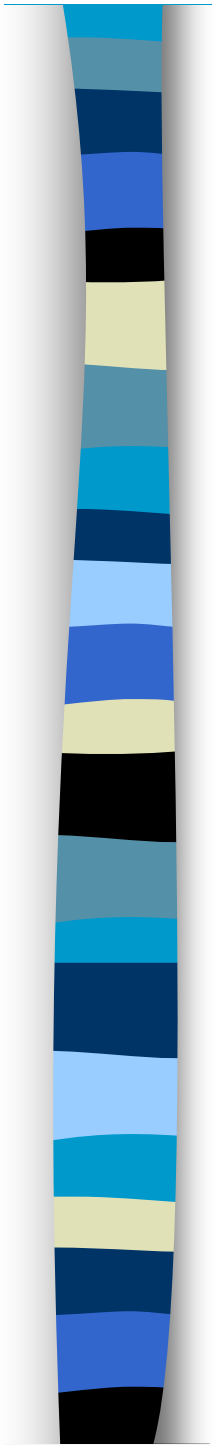


– E.g. 6 **Constrained** dependent variable
  - E.g. Loan to value ratios
  - E.g. % unemployed
  - Solution: use Fractional Logit Regression



– E.g. 7 **Grouped Data:** occurs when we have apparently ordered data but where the threshold values for categories are known:
  - e.g. a survey of incomes, which is coded as follows:
    = 1 if income < 5,000,
    = 2 if 5,000 ≤ income < 7,000,
    = 3 if 7,000 ≤ income < 10,000,
    = 4 if 10,000 ≤ income < 15,000,
    = 5 if income ≥ 15,000
  - *If the Dependent variable is Censored, Estimate using:* Grouped Tobit (e.g. LIMDEP)

- Ambiguity:
  - The level of measurement of a variable is sometimes ambiguous:
    - "...statements about levels of measurement of a [variable] cannot be sensibly made in isolation from the theoretical and substantive context in which the [variable] is to be used" (Carter, 1971, p.12, quoted in Long 1997, p. 2)
  - e.g. education: could be measured as a:
    - » *binary variable*: 1 if only attained High School or less, 0 if other.
    - » *ordinal variable*: coded 6 if has PhD, 5 if has Masters, 4 if has Degree, 3 if has Highers, 2 if has Standard Grades, 1 if no qualifications
    - » *count variable*: number of school years completed

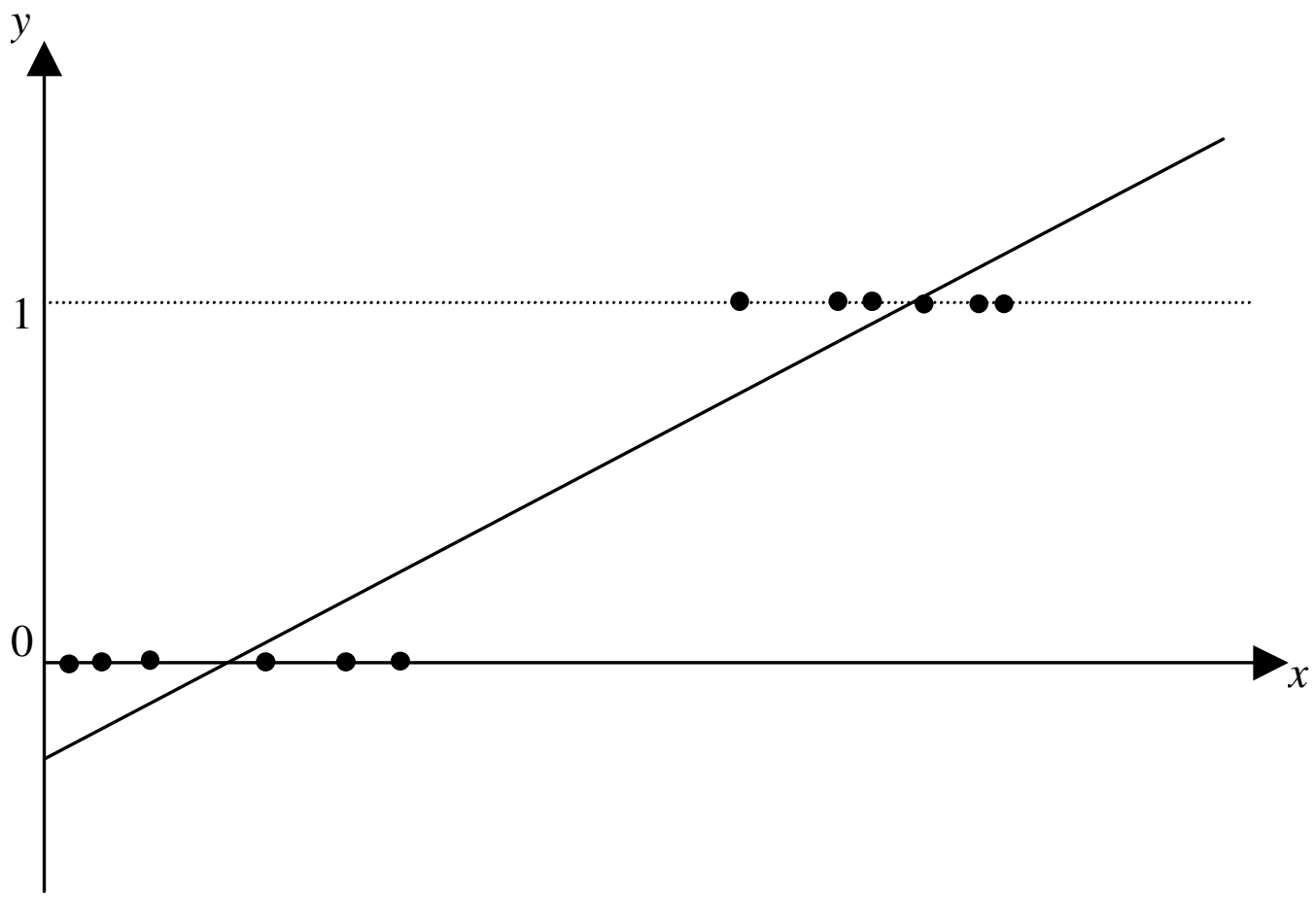


- Choosing the Appropriate Statistical Models:
  - if we choose a model that assumes a level of measurement of the dependent variable different to that of our data, then the estimates may be:
    - biased,
    - inefficient
    - or inappropriate
  - e.g. if we apply standard OLS to dependent variables that fall into any of the above categories of data, it will assume that the variable is unbounded and continuous and construct a line of best fit accordingly
- In this lecture we shall only look at the **logit** model

2. Linear Probability Model

- Q/ What happens if we try to fit a line of best fit to a regression where the dependent variable is binary?
  - Draw a scatter plot
  - draw a line of best fit
  - what is the main problem with the line of best fit?
  - How might a correct line of best fit look?

Linear Probability Model:

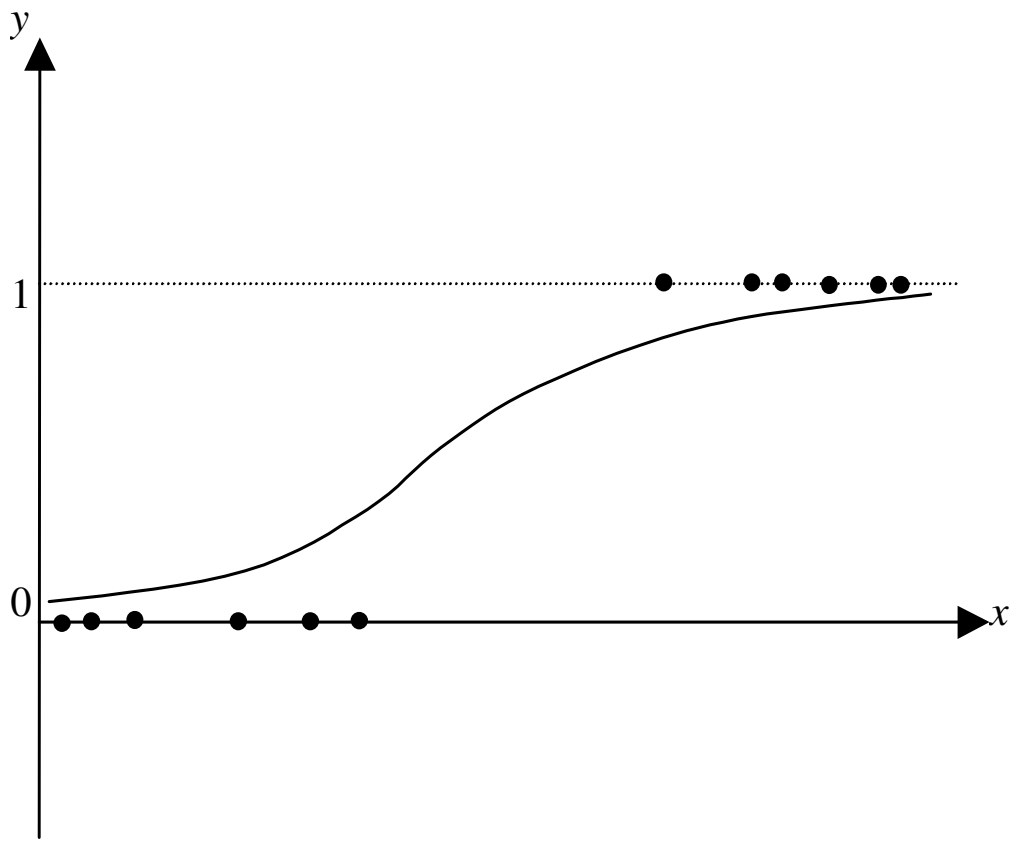


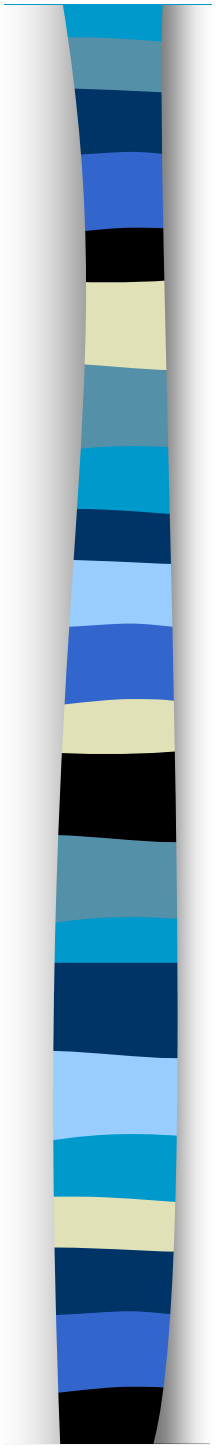
- Advantage:
  - interpretation is straightforward:
    - the coefficient is interpreted in the same way as linear regression
      - e.g. Predicted Probability of Labour Force Participation
        - » if b_1 = 0.4, then the predicted probability of labour force participation increases by 0.4, holding all other variables constant.

- Disadvantages:
  - *heteroscedasticity:*
    - error term will tend to be larger for middle values of *x*
    - OLS estimates are inefficient and standard errors are biased, resulting in incorrect t-statistics.
  - *Non-normal errors:*
    - but normality not required for OLS to be BLUE
  - *Nonsensical Predictions*:
    - Predicted values can be < 0, or > 1.

- *Functional Form:*
  - the nonsensical predictions arise because we are trying to fit a linear function to a fundamentally non-linear relationship:
    - *probabilities tend to have a non-linear relationship with their determinants:*
    - e.g. cannot say that each additional child will remove 0.4 from the probability of labour force participation:
      - » Prob(LF particip. of 20 year old Female with no children) = 0.5
      - » Prob(LF particip. of 20 year old Female with 1 child) = 0.1
      - » Prob(LF particip. of 20 year old Female with 2 children) = -0.3

True functional form:





- Q/ What kind of model/transformation of our data could be used to represent this kind of relationship?
  - I.e. one that is:
    - "*s*" shaped
    - converges to zero at one end and converges to 1 at the other end
      - this rules out cubic transformations since they are unbounded

- Note also that we may well have more than one explanatory variable, so we need a model that can transform:

  $b_0 + b_1x_1 + b_2x_2 + b_3x_3$

  into values for *y* that range between 0 and 1

3. Logit:

- One popular transformation is the logit or logistic transformation:
  - Can apply this to a single variable:

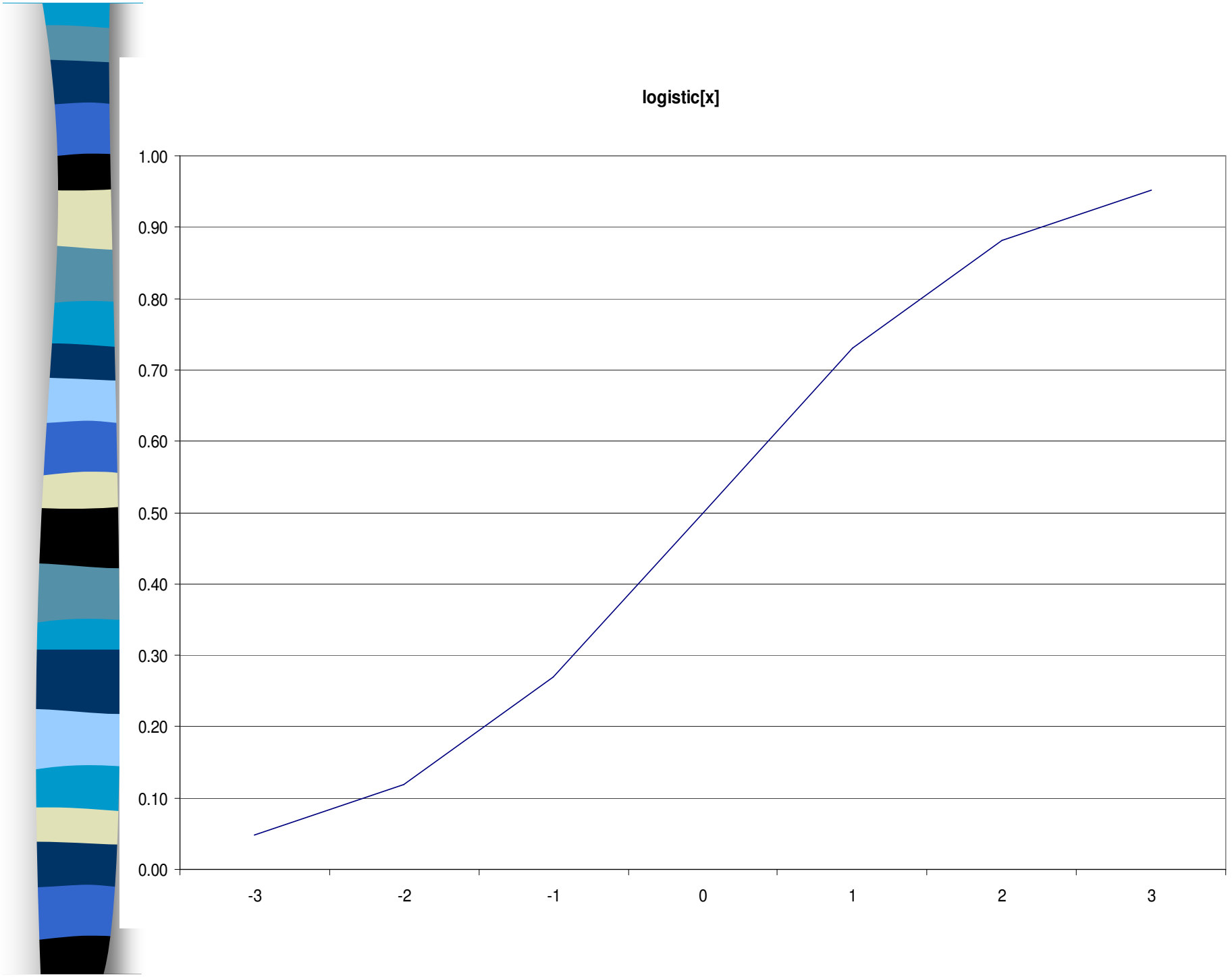
$$\frac{\exp(x)}{1+\exp(x)}$$

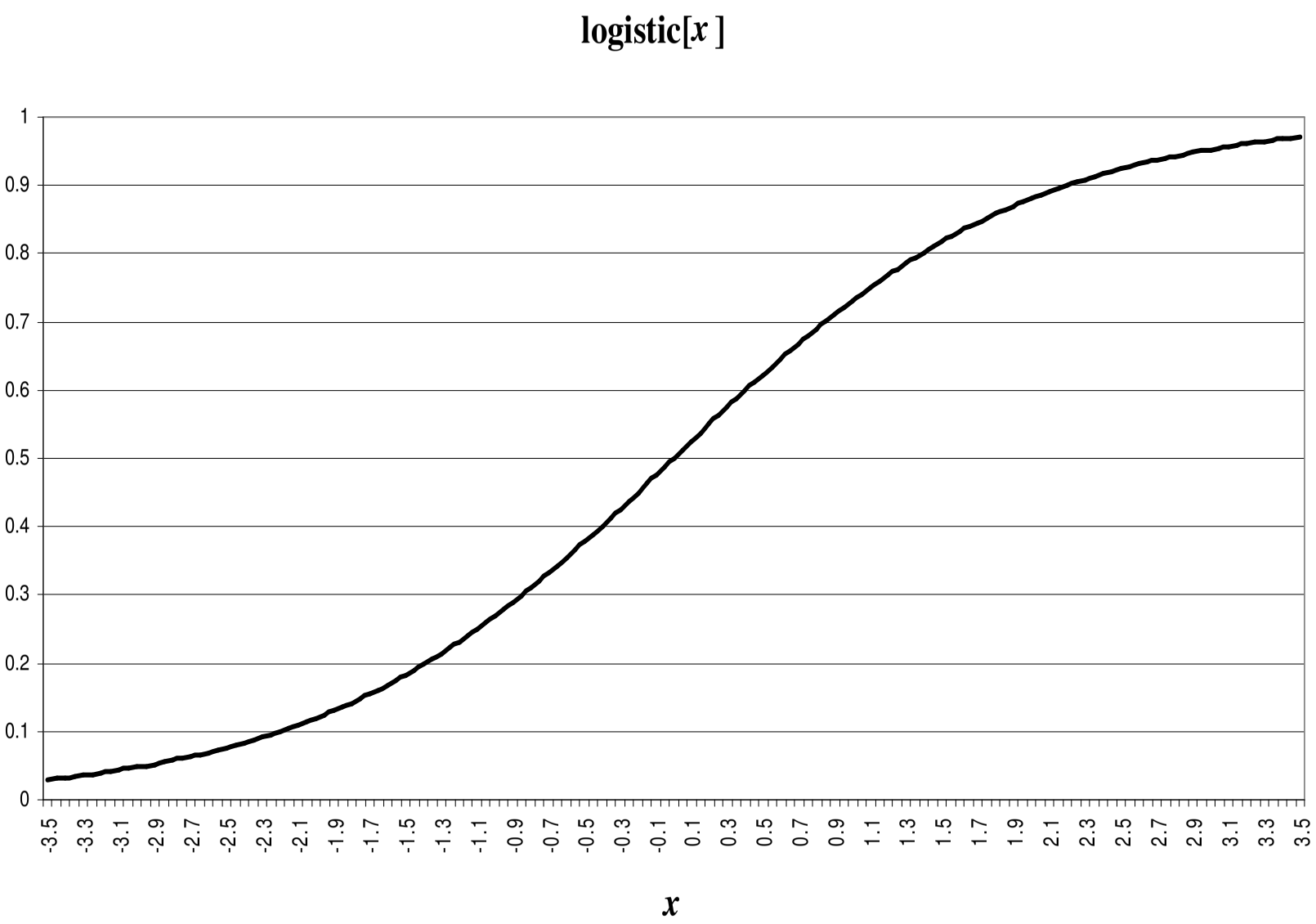
- Or apply to an entire function

$$\frac{\exp(b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3)}{1+\exp(b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3)}$$

E.g. Calculation for Logistic Distribution

x	exp(x)	1+exp(x)	logit[x]
-3	0.05	1.05	0.05
-2	0.14	1.14	0.12
-1	0.37	1.37	0.27
0	1.00	2.00	0.50
1	2.72	3.72	0.73
2	7.39	8.39	0.88
3	20.09	21.09	0.95


logistic[x]
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
-3
-2
-1
0
1
2
3

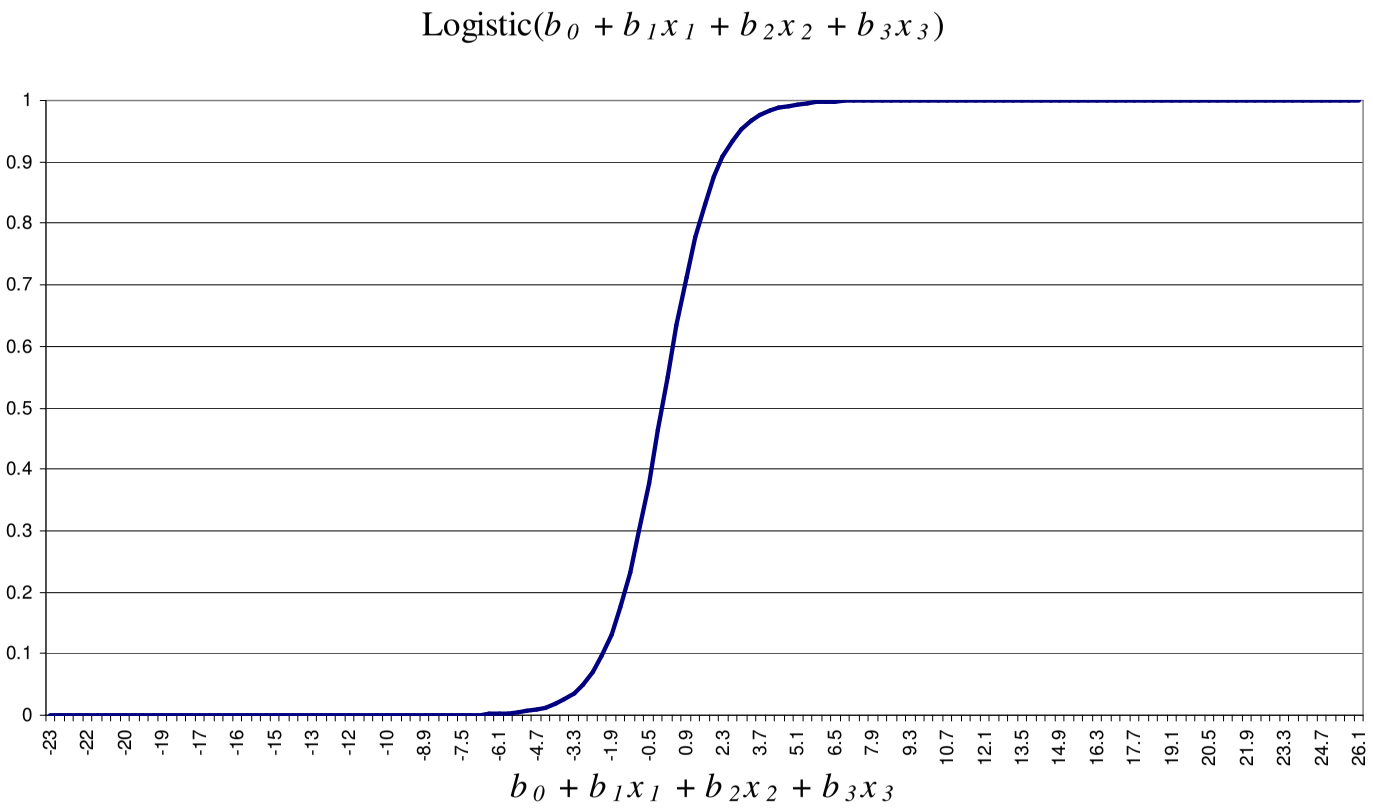

logistic[x]
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-3.5
-3.3
-3.1
-2.9
-2.7
-2.5
-2.3
-2.1
-1.9
-1.7
-1.5
-1.3
-1.1
-0.9
-0.7
-0.5
-0.3
-0.1
0.1
0.3
0.5
0.7
0.9
1.1
1.3
1.5
1.7
1.9
2.1
2.3
2.5
2.7
2.9
3.1
3.3
3.5
x

More than one explanatory variable:

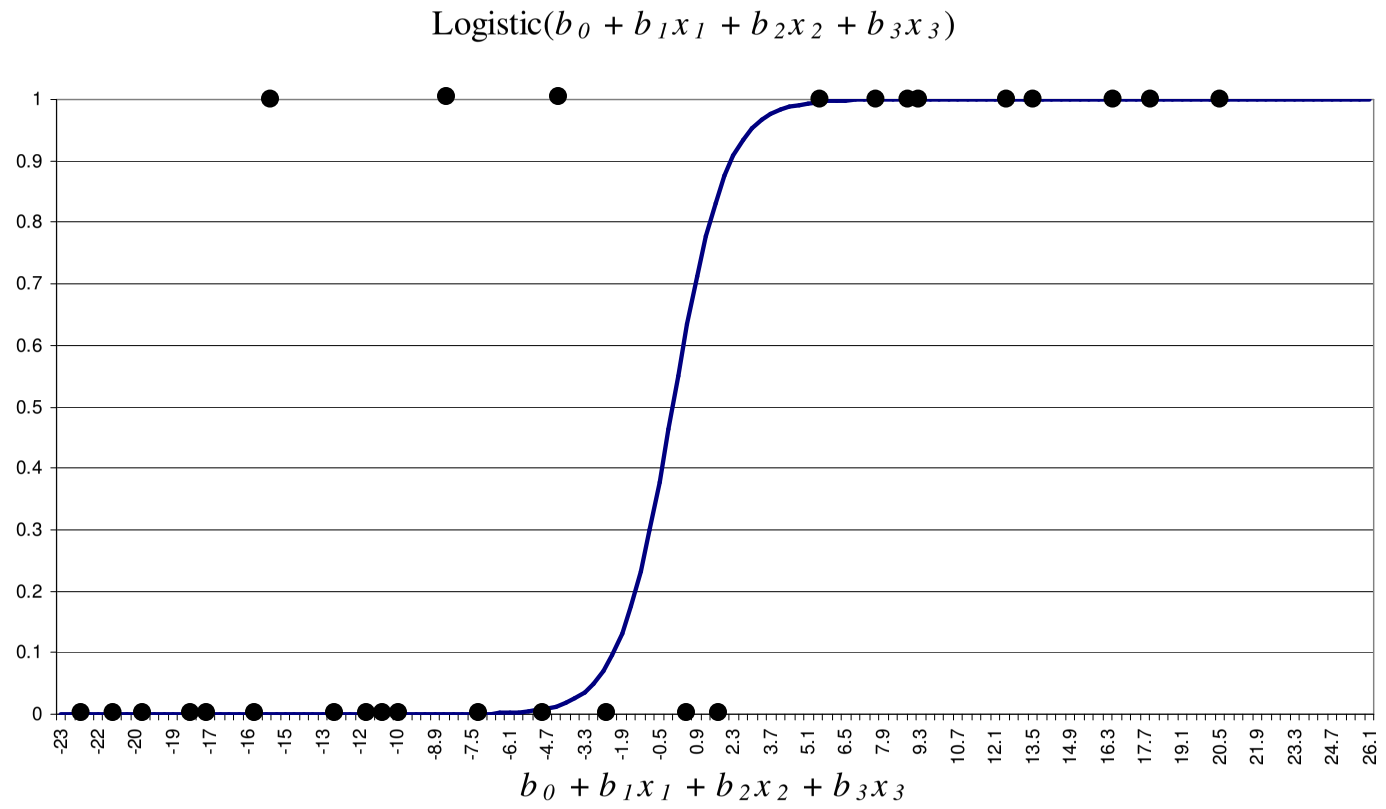
$b_0 + b_1x_1 + b_2x_2 + b_3x_3 =$ $230 - 4x_1 + 7x_2 + 8x_3$

b0	x1	x2	x3	b1x1 + b2x2 + b3x3	Logistic()
230	0	-0.7	-31	-22.9	1.13411E-10
230	1	-0.65	-30.5	-22.55	1.60938E-10
230	2	-0.6	-30	-22.2	2.28382E-10
230	3	-0.55	-29.5	-21.85	3.2409E-10
230	4	-0.5	-29	-21.5	4.59906E-10
230	5	-0.45	-28.5	-21.15	6.52637E-10
230	6	-0.4	-28	-20.8	9.26136E-10
230	7	-0.35	-27.5	-20.45	1.31425E-09
230	8	-0.3	-27	-20.1	1.86501E-09
230	9	-0.25	-26.5	-19.75	2.64657E-09
230	10	-0.2	-26	-19.4	3.75567E-09
230	11	-0.15	-25.5	-19.05	5.32954E-09
230	12	-0.1	-25	-18.7	7.56298E-09
230	13	-0.05	-24.5	-18.35	1.07324E-08
230	14	9.71445E-17	-24	-18	1.523E-08
230	15	0.05	-23.5	-17.65	2.16124E-08
230	16	0.1	-23	-17.3	3.06694E-08
230	17	0.15	-22.5	-16.95	4.3522E-08

Plot for full range of values of the *x*'s:



Observed values of *y* included:



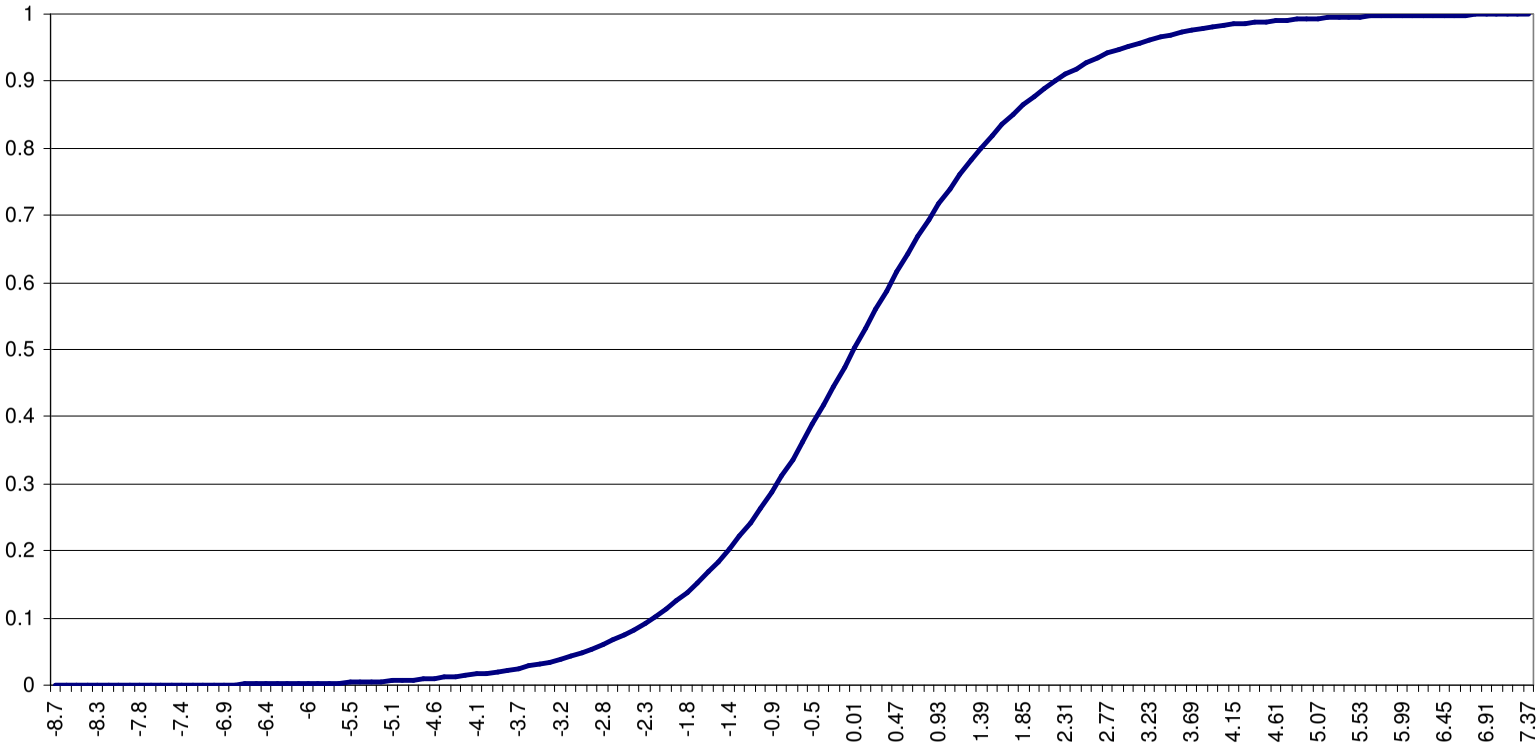
- Goodness of fit:
  - if observed values of y were were found for a wide range of the possible values of *x*, then this plot wouldn't be a very good line of best fit
  - values of $b_0 + b_1x_1 + b_2x_2 + b_3x_3$ that are less than -4 or greater than 4 have very little effect on the probability
  - yet most of the values of *x* lie outside the -4, 4 range.
  - Perhaps if we alter the estimated values of b_k then we might improve our line of best fit...

Suppose we try:

$b_0 = 22$, $b_1 = -0.4$, $b_2 = 0.5$ and $b_3 = 0.98$:

b0	x1	x2	x3	22 -0.4x1 + 0.5x2 + 0.98x3	Logistic()
22	0	-0.7	-31	-8.73	0.000162
22	1	-0.65	-30.5	-8.615	0.000181
22	2	-0.6	-30	-8.5	0.000203
22	3	-0.55	-29.5	-8.385	0.000228
22	4	-0.5	-29	-8.27	0.000256
22	5	-0.45	-28.5	-8.155	0.000287
22	6	-0.4	-28	-8.04	0.000322
22	7	-0.35	-27.5	-7.925	0.000361
22	8	-0.3	-27	-7.81	0.000405
22	9	-0.25	-26.5	-7.695	0.000455
22	10	-0.2	-26	-7.58	0.00051
22	11	-0.15	-25.5	-7.465	0.000572
22	12	-0.1	-25	-7.35	0.000642
22	13	-0.05	-24.5	-7.235	0.00072
22	14	9.71E-17	-24	-7.12	0.000808
22	15	0.05	-23.5	-7.005	0.000907
22	16	0.1	-23	-6.89	0.001017
22	17	0.15	-22.5	-6.775	0.001141
22	18	0.2	-22	-6.66	0.00128
22	19	0.25	-21.5	-6.545	0.001435

$$\text{Logistic}(b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3)$$



$$22 - 0.4x_1 + 0.5x_2 + 0.98x_3$$

4. Logit Estimation & Interpretation

- The above discussion leads naturally to a probability model of the form:

$$\Pr(y=1 \mid x_k) = \frac{\exp(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_K x_K)}{1 + \exp(b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_K x_K)}$$

- We now need to find a way of estimating values of b_k that will best fit the data.
- Unfortunately, OLS cannot be applied since the above model is non-linear in parameters.

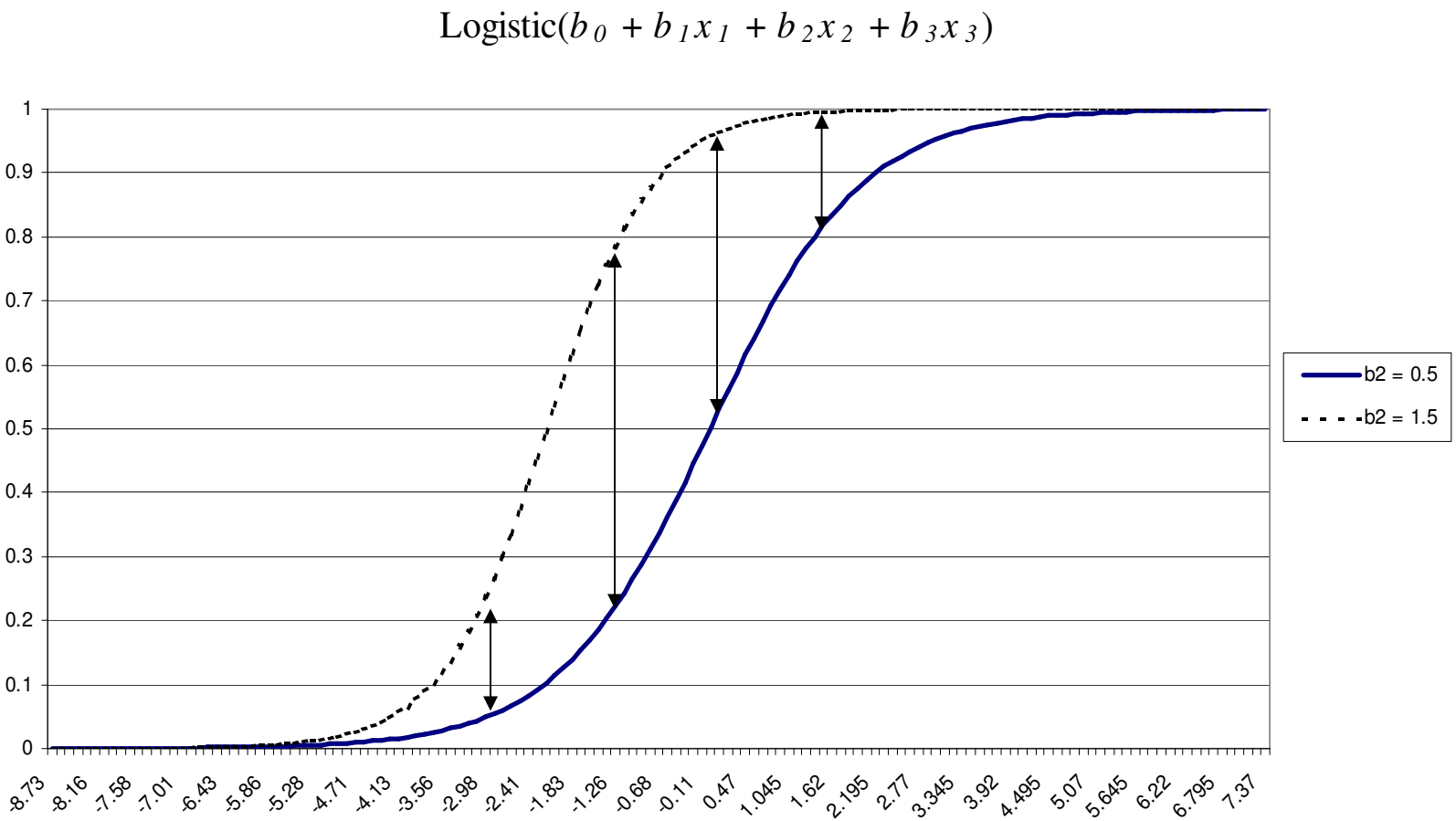
Maximum Likelihood

- The method used to estimate logit is maximum likelihood:
  - starts by saying, for a given set of parameter values, what is the probability of observing the current sample.
  - It then tries various values of the parameters to arrive at estimates of the parameters that makes the observed data most likely

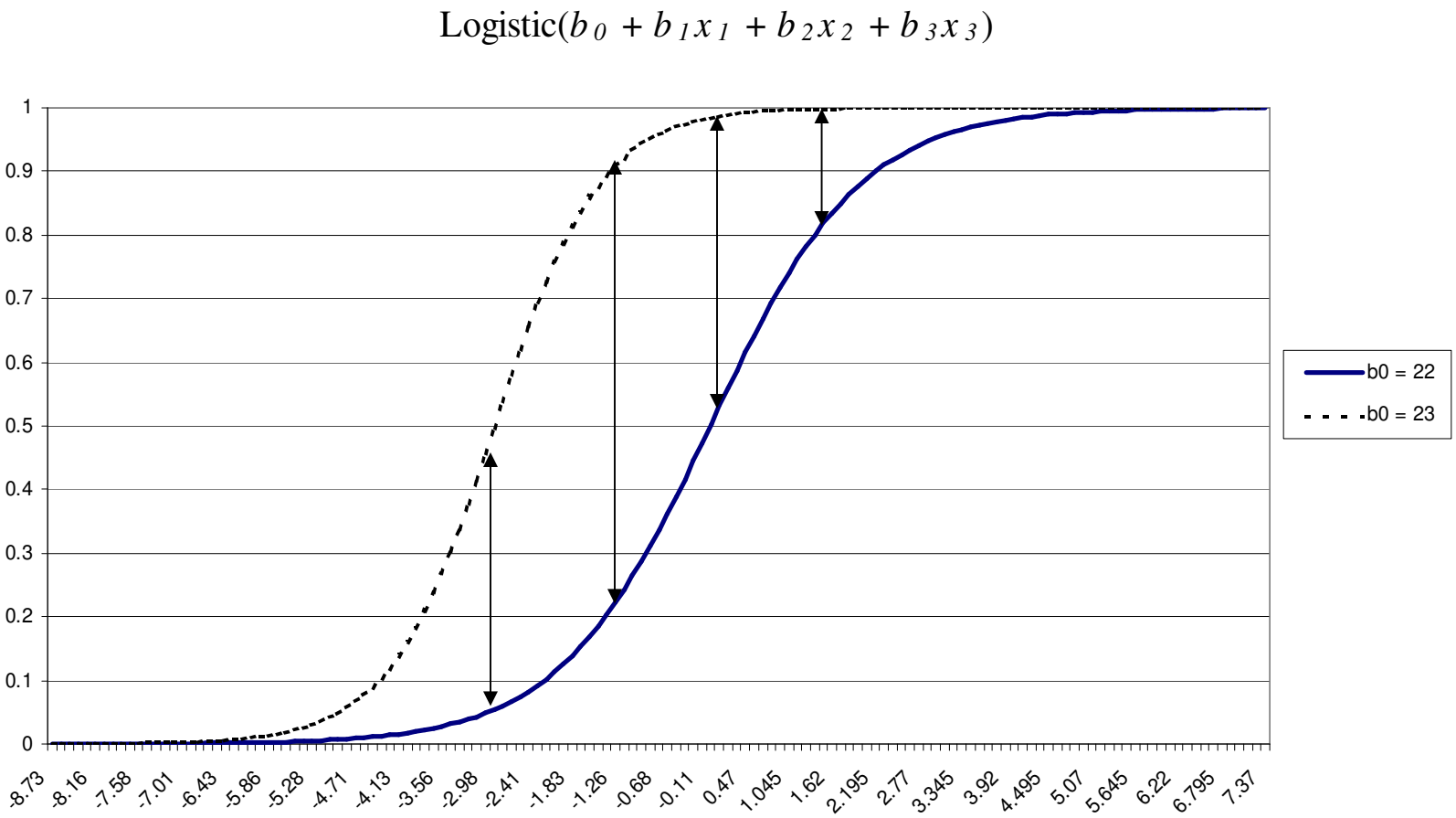
Interpreting the Logit Output

- Because logit regression is fundamentally non-linear, interpretation of output can be difficult
- many studies that use logit overlook this fact:
  - either interpret magnitude of coefficients incorrectly
  - or only interpret signs of coefficients

Impact of increasing b_2 by 1:



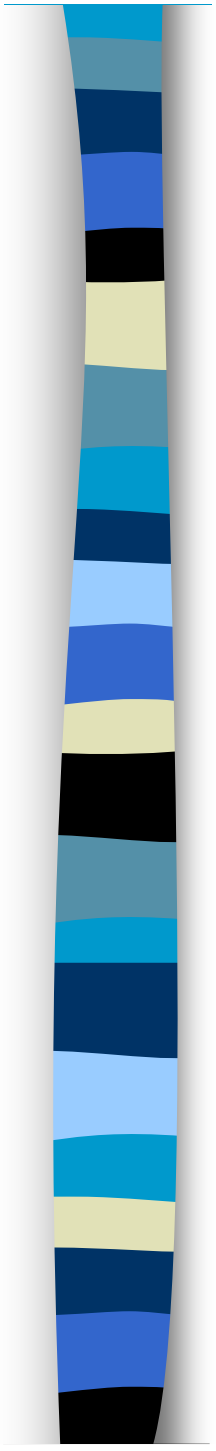
Impact of increasing b_0 by 1:



E.g. Using Logit to Estimate the Effect of No. Children on MPPI take-up

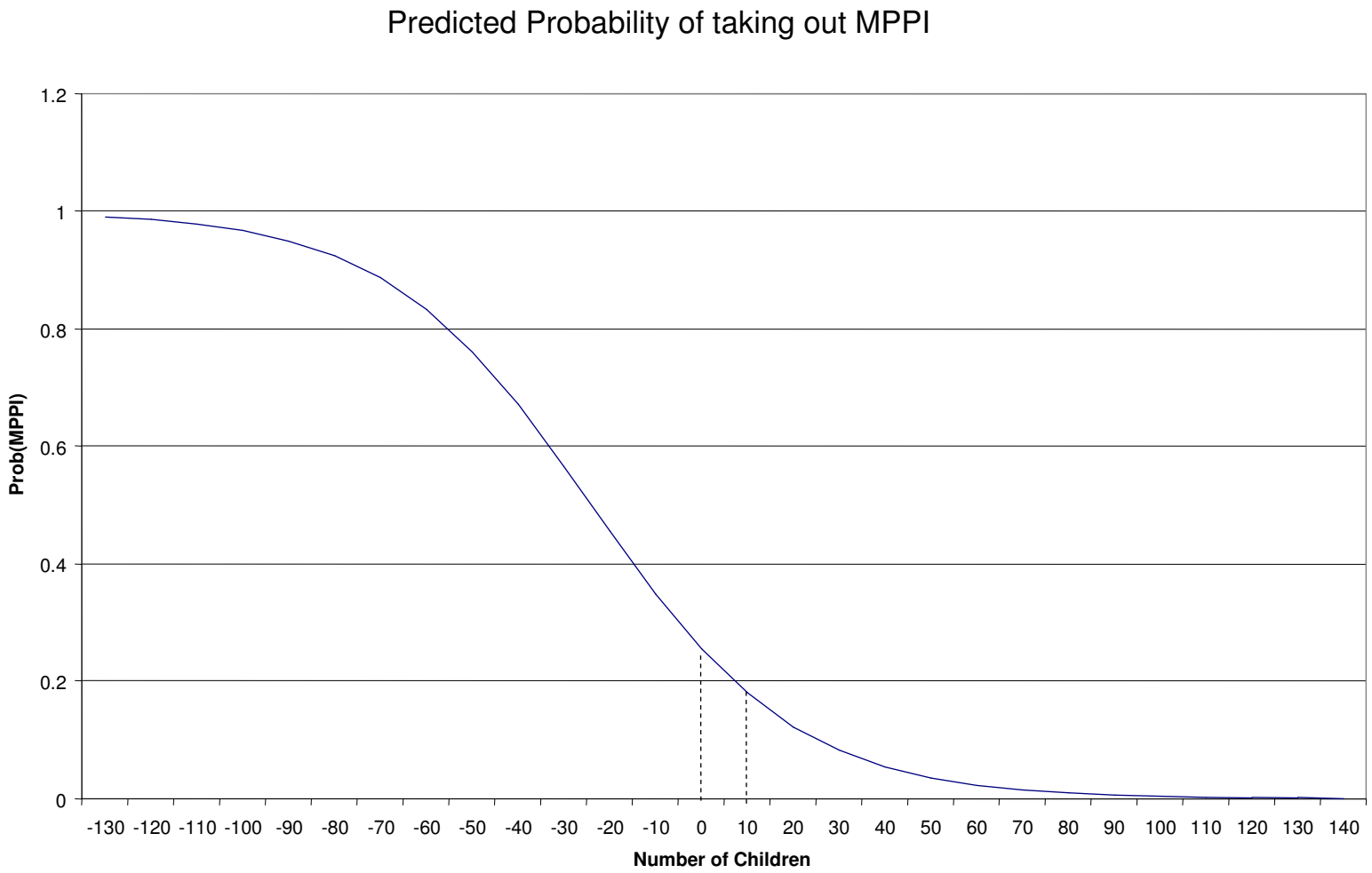
---------- Variables in the Equation ------

Variable	B	S.E.	Wald	df	Sig
CHILDREN	-.0446	.0935	.2278	1	.6331
Constant	-1.0711	.1143	87.8056	1	.0000



b0	b1	x1	b0 + b1x1	Predicted Probability of taking out MPPI
-1.0711	-0.0446	-130	4.7269	0.991223828
-1.0711	-0.0446	-120	4.2809	0.986358456
-1.0711	-0.0446	-110	3.8349	0.978853343
-1.0711	-0.0446	-100	3.3889	0.967355826
-1.0711	-0.0446	-90	2.9429	0.949926848
-1.0711	-0.0446	-80	2.4969	0.923924213
-1.0711	-0.0446	-70	2.0509	0.886038527
-1.0711	-0.0446	-60	1.6049	0.832702114
-1.0711	-0.0446	-50	1.1589	0.761132782
-1.0711	-0.0446	-40	0.7129	0.671041637
-1.0711	-0.0446	-30	0.2669	0.566331702
-1.0711	-0.0446	-20	-0.1791	0.455344304
-1.0711	-0.0446	-10	-0.6251	0.348622427
-1.0711	-0.0446	0	-1.0711	0.255193951
-1.0711	-0.0446	10	-1.5171	0.179888956
-1.0711	-0.0446	20	-1.9631	0.123131948
-1.0711	-0.0446	30	-2.4091	0.082481403
-1.0711	-0.0446	40	-2.8551	0.05441829
-1.0711	-0.0446	50	-3.3011	0.035533472

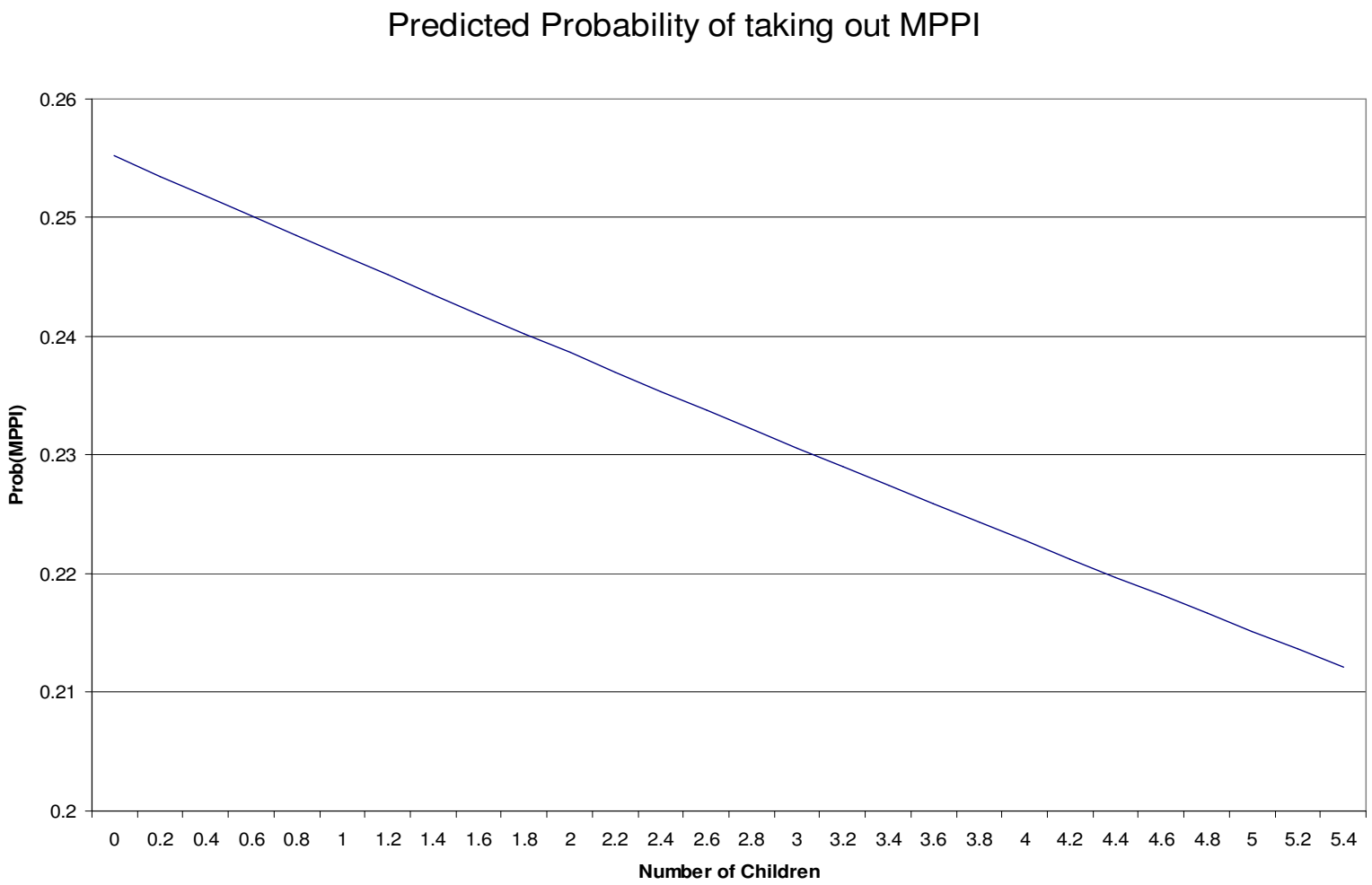
Predicted values:



Predicted probabilities over the relevant range of values of *x*:

b0	b1	x1		Predicted Probability of taking out MPPI
-1.0711	-0.0446	0	-1.0711	0.255193951
-1.0711	-0.0446	1	-1.1157	0.24680976
-1.0711	-0.0446	2	-1.1603	0.238612778
-1.0711	-0.0446	3	-1.2049	0.230604681
-1.0711	-0.0446	4	-1.2495	0.222786703
-1.0711	-0.0446	5	-1.2941	0.215159652
-1.0711	-0.0446	6	-1.3387	0.207723924
-1.0711	-0.0446	7	-1.3833	0.200479528
-1.0711	-0.0446	8	-1.4279	0.193426099

Predicted values over relevant values of *x*:



5. Multiple Logit Regression:

- More complex if have more than one *x* since the effect on the dependent variable will depend on the values of the other explanatory variables.
- One solution to this is to use the *odds*:

$$\text{odds} = \frac{P(\text{event})}{P(\text{no event})} = \frac{P(\text{event})}{1 - P(\text{event})}$$

Computing odds:

- Compute the odds associated probabilities:
  1. The probability of Wales winning the 6 Nations rugby tournament = 60%, odds = _____?
  2. The probability of England winning the 6 Nations = 50%, odds = _____?
  3. The probability of MPPI take-up in a HH with no children = 25%, odds = _____?
  4. Pr(MPPI take-up | one child) = .24, odds = _____?
  5. Pr(passing SSS2) = 0.9, odds = _____?

- Compute the odds associated probabilities:
  1. The probability of Wales winning the 6 Nations rugby tournament = 60%, odds = 6/4 = 1.5
  2. The probability of England winning the 6 Nations = 50%, odds = 1/1 = 1.
  3. The probability of MPPI take-up in a HH with no children = 25%, odds = 1/3 = .33 .
  4. Pr(MPPI take-up | one child) = .24, odds = .24/.76 = .32
  5. Pr(passing SSS2) = 0.9, odds = 9/1 = 9.

Probability	Odds
0.60	1.50
0.50	1.00
0.25	0.33
0.24	0.32
0.90	9.00

- SPSS calculates "**Exp(B)**" which gives the *proportionate change in the predicted odds* of taking out MPPI due to a unit change in the explanatory variable, holding all other variables constant:

Variable	B	S.E.	Exp(B)
CHILDREN	-.0446	.0935	.9564
Constant	-1.0711	.1143	

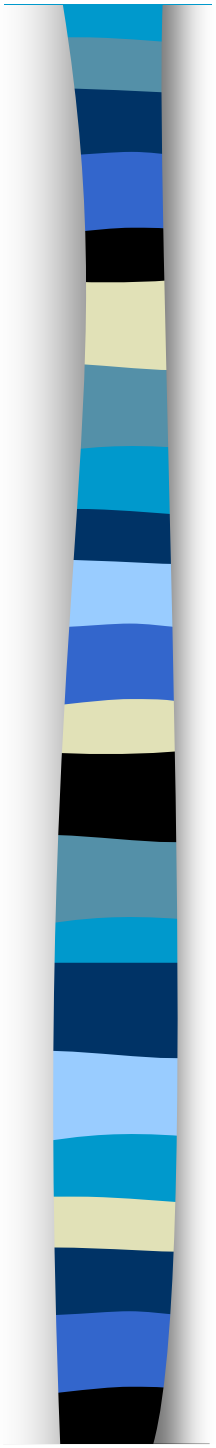
b0	b1	x1		Predicted Probability of taking out MPPI
-1.0711	-0.0446	0	-1.0711	0.255193951
-1.0711	-0.0446	1	-1.1157	0.24680976

- E.g. effect on the predicted odds of taking out MPPI of having 1 more child:
- Prob(MPPI|child = 0) = 0.2552
- Odds(MPPI|child = 0) = 0.2552/(1-0.2552) = **0.3426**

- Prob(MPPI|child = 1) = 0.2468
- Odds(MPPI|child = 1) = 0.2468/(1-0.2468) = **0.3277**

- Proport.Change in Odds = odds after a unit change in the predictor / original odds

  = **Exp(B)** = 0.3277 / 0.3426 = **0.956**



- Notes:
  - if the value of **`Exp(B)`** is > 1 then it indicates that as the explanatory variable increases, the odds of the outcome occurring increase.
  - if the value of **`Exp(B)`** is < 1 then it indicates that as the explanatory variable increases, the odds of the outcome occurring decrease.
    - I.e. between zero and 1

Reading:

- Kennedy, P. “A Guide to Econometrics” chapter 15
- Field, A. “Discovering Statistics”, chapter 5.

For a more comprehensive treatment of this topic, you may want to consider purchasing:

- Scott, J. S.(1997) “Regression models for Categorical and Limited Dependent Variables”, Sage: Thousand Oaks California.

  This is a technical but first rate introduction to logit -- thorough but clear -- well worth purchasing if you are going to do any amount of work using logit, probit or any other qualitative response model. Probably the best book around on the subject.