NCRM NATIONAL CENTRE FOR RESEARCH METHODS

Statistical Modelling Workshop (October 2022)

A one day training workshop led by

Professor Vernon Gayle
University of Edinburgh

Many of the ideas and examples presented in this file draw heavily on
previous work. We are grateful for the comments and feedback from participants
of earlier workshops.

Professor Vernon Gayle (vernon.gayle@ed.ac.uk)

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

Statistical Modelling

Observational Studies

- Social surveys
- Census records
- Administrative data records
- Digital records

Experimental Data

- Explicit Design (random allocation)
- H_o Group A = Group B
- Small *n*
- Few variables *k*

Observational Data

- General data collection
- Large *n*
- Many variables *k* (including key variables)

Data Editor (Browse) - [a_indall]

File Edit View Data Tools

pidp[1] 68001367

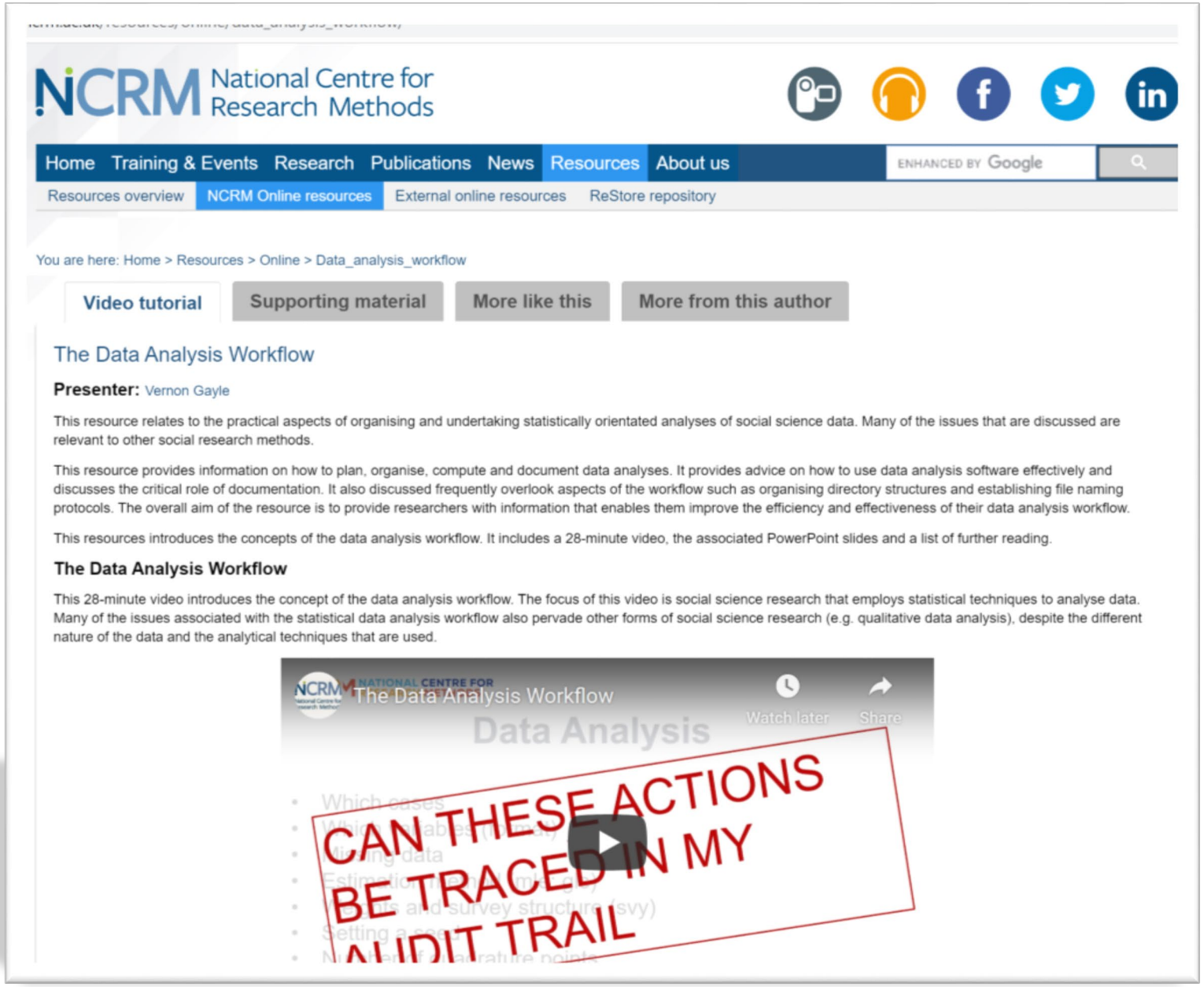
	pidp	a_hidp	a_pno	a_hhorig	a_memorig	a_psu	a_strata	a_sampst	a_month	a_iv
1	68001367	68001363	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
2	68004087	68004083	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
3	68006127	68006123	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
4	68006131	68006123	2	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	refusal
5	68006135	68006123	3	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
6	68006139	68006123	4	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	youth interview
7	68006807	68006803	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
8	68007487	68007483	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
9	68007491	68007483	2	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	proxy interview
10	68007495	68007483	3	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	proxy interview
11	68007499	68007483	4	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	proxy interview
12	68008167	68008163	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
13	68008171	68008163	2	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	proxy interview
14	68008847	68008843	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
15	68009527	68009523	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
16	68010207	68010203	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
17	68010887	68010883	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
18	68010891	68010883	2	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	proxy interview
19	68011567	68011563	1	ukhls gb 2009-10	ukhls gb 2009-10	2012	2006	osm	jan yr1	full interview
20	68014287	68014283	1	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	full interview
21	68014291	68014283	2	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	full interview
22	68014295	68014283	3	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	child under 10
23	68014299	68014283	4	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	child under 10
24	68014967	68014963	1	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	full interview
25	68016327	68016323	1	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	full interview
26	68017687	68017683	1	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	full interview
27	68017691	68017683	2	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	other non-intvw
28	68017695	68017683	3	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	child under 10
29	68017699	68017683	4	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	child under 10
30	68019047	68019043	1	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	full interview
31	68019051	68019043	2	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	full interview
32	68019055	68019043	3	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	child under 10
33	68020407	68020403	1	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	full interview
34	68023127	68023123	1	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	full interview
35	68023131	68023123	2	ukhls gb 2009-10	ukhls gb 2009-10	2036	2018	osm	jan yr1	full interview
36	68025167	68025163	1	ukhls gb 2009-10	ukhls gb 2009-10	2060	2030	osm	jan yr1	full interview
37	68025847	68025843	1	ukhls gb 2009-10	ukhls gb 2009-10	2060	2030	osm	jan yr1	full interview

Variables Snapshots Properties

Ready

Vars: 131 Order: Dataset Obs: 77,309 Filter: Off Mode: Browse CAP NUM

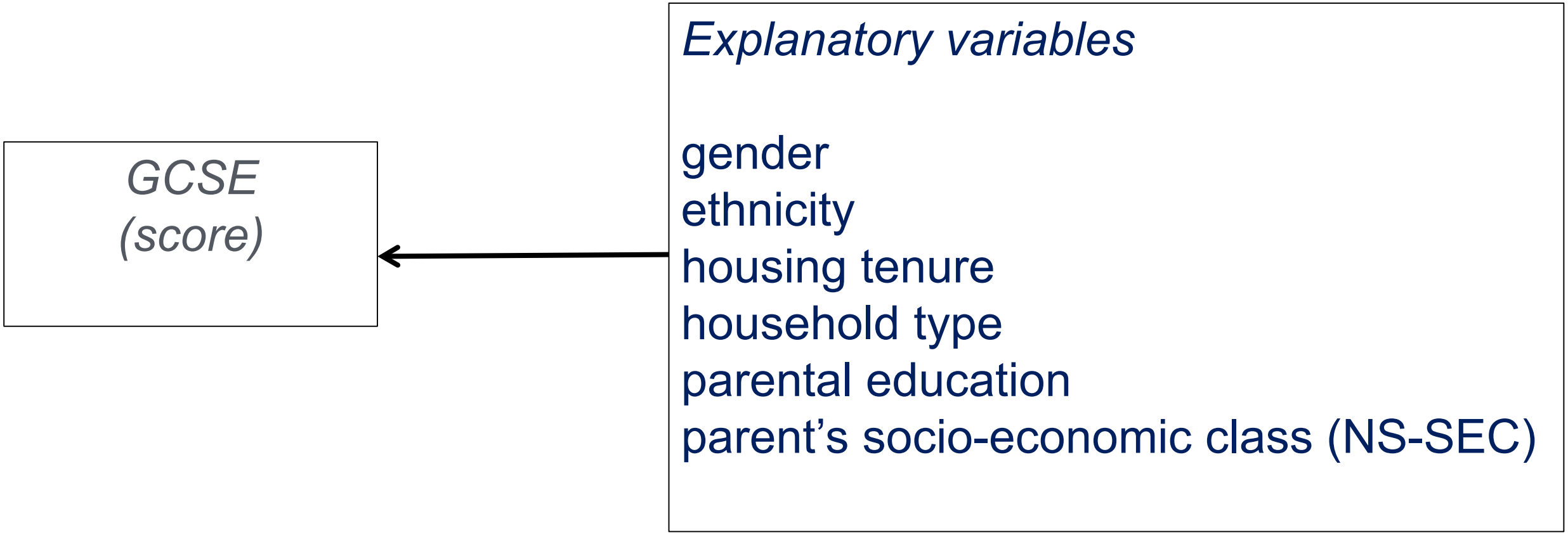

Excel



https://www.ncrm.ac.uk/resources/online/data_analysis_workflow/

Sophisticated Description

Gayle, V., Murray, S. and Connelly, R., 2016. Young people and school General Certificate of Secondary Education attainment: Looking for the 'missing middle'. *British Journal of Sociology of Education*, *37*(3), pp.350-370.



https://www.tandfonline.com/doi/pdf/10.1080/01425692.2014.935292
http://bit.do/gayle_et_2016

Generalize Linear Models (glm)

J. R. Statist. Soc. A,
(1972), **135**, *Part* 3, p. 370

Generalized Linear Models

By J. A. NELDER and R. W. M. WEDDERBURN

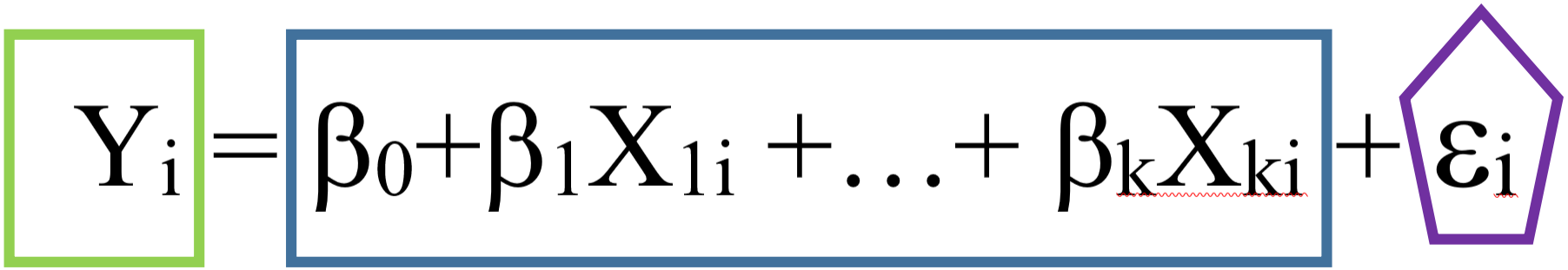
Rothamsted Experimental Station, Harpenden, Herts

SUMMARY

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

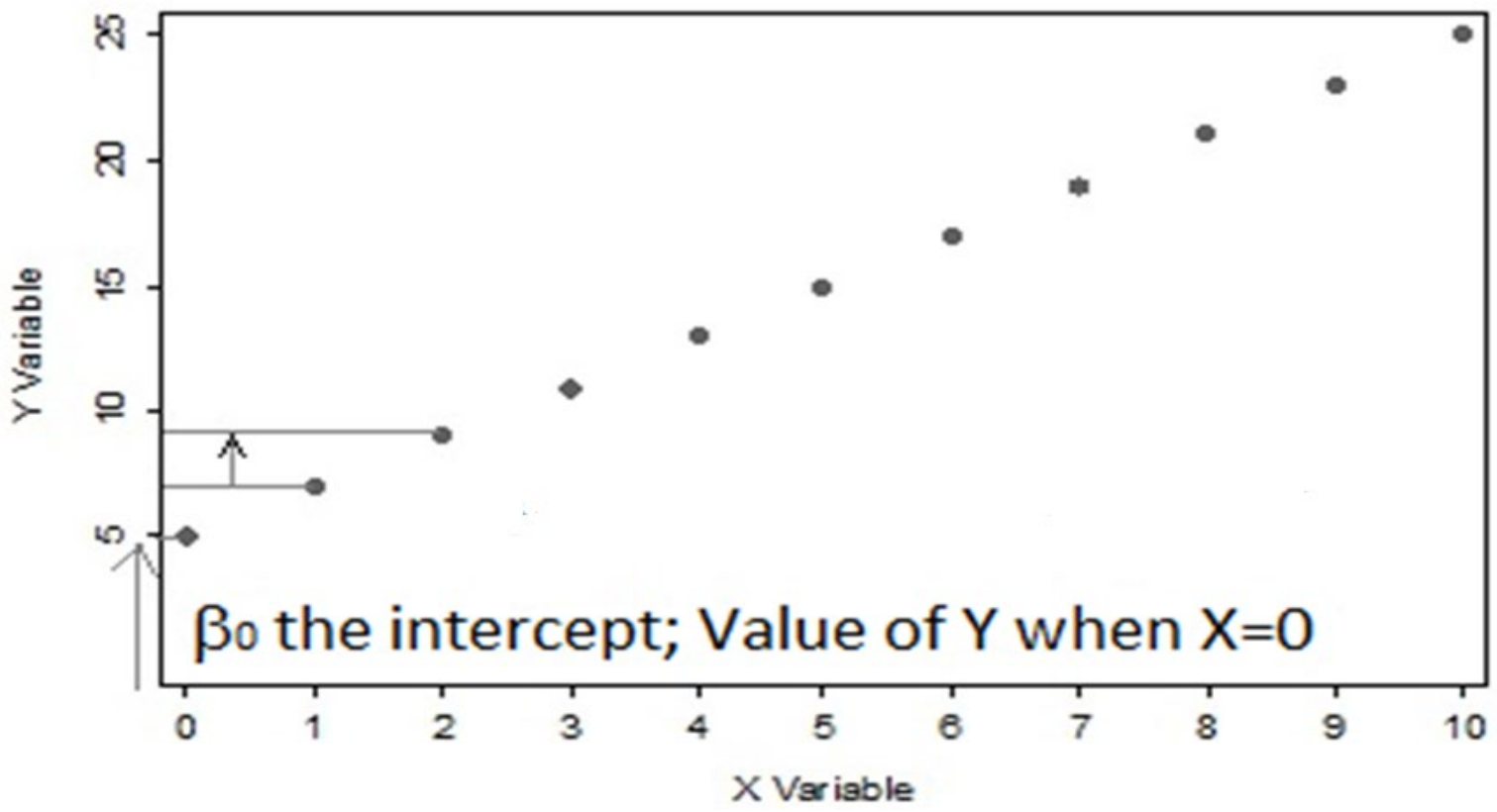
Keywords: ANALYSIS OF VARIANCE; CONTINGENCY TABLES; EXPONENTIAL FAMILIES; INVERSE POLYNOMIALS; LINEAR MODELS; MAXIMUM LIKELIHOOD; QUANTAL RESPONSE; REGRESSION; VARIANCE COMPONENTS; WEIGHTED LEAST SQUARES


$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

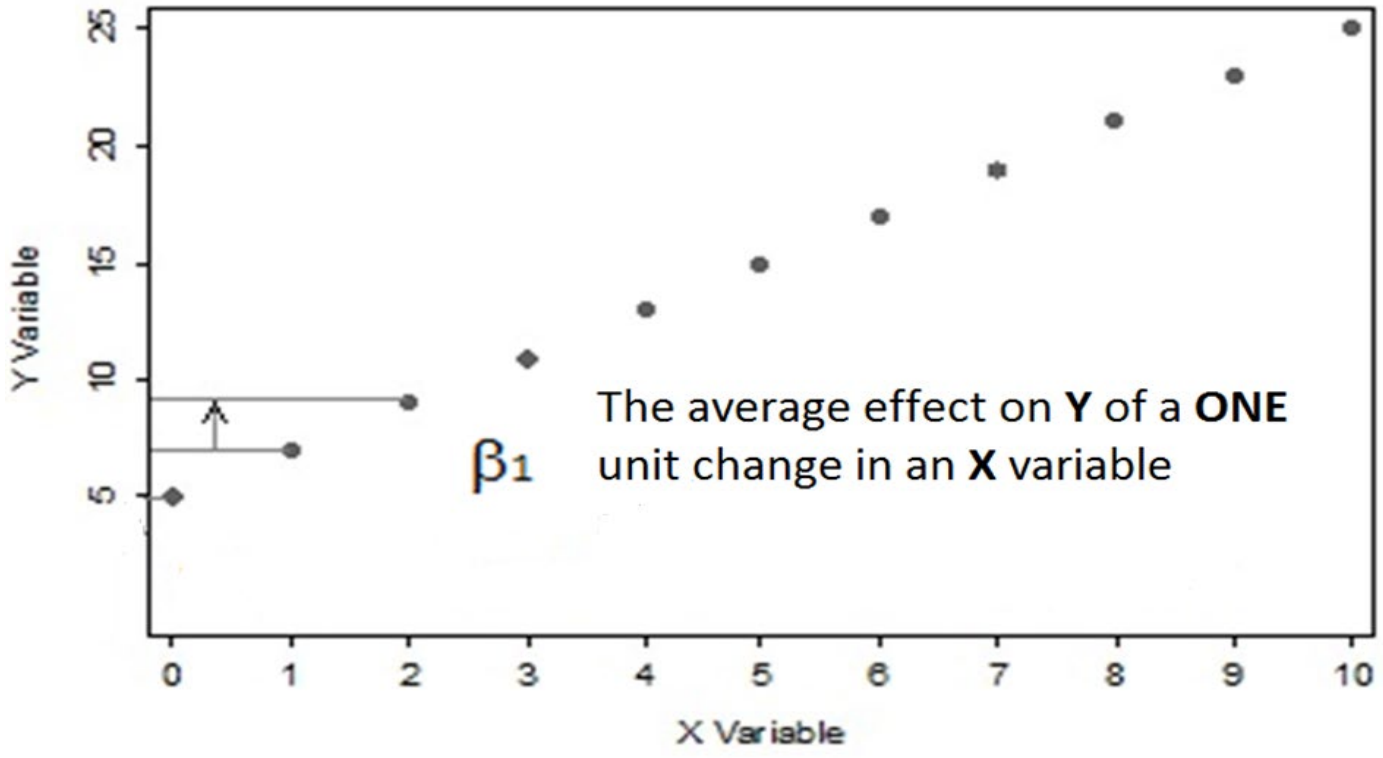
$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

The Intercept (β_o)



$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

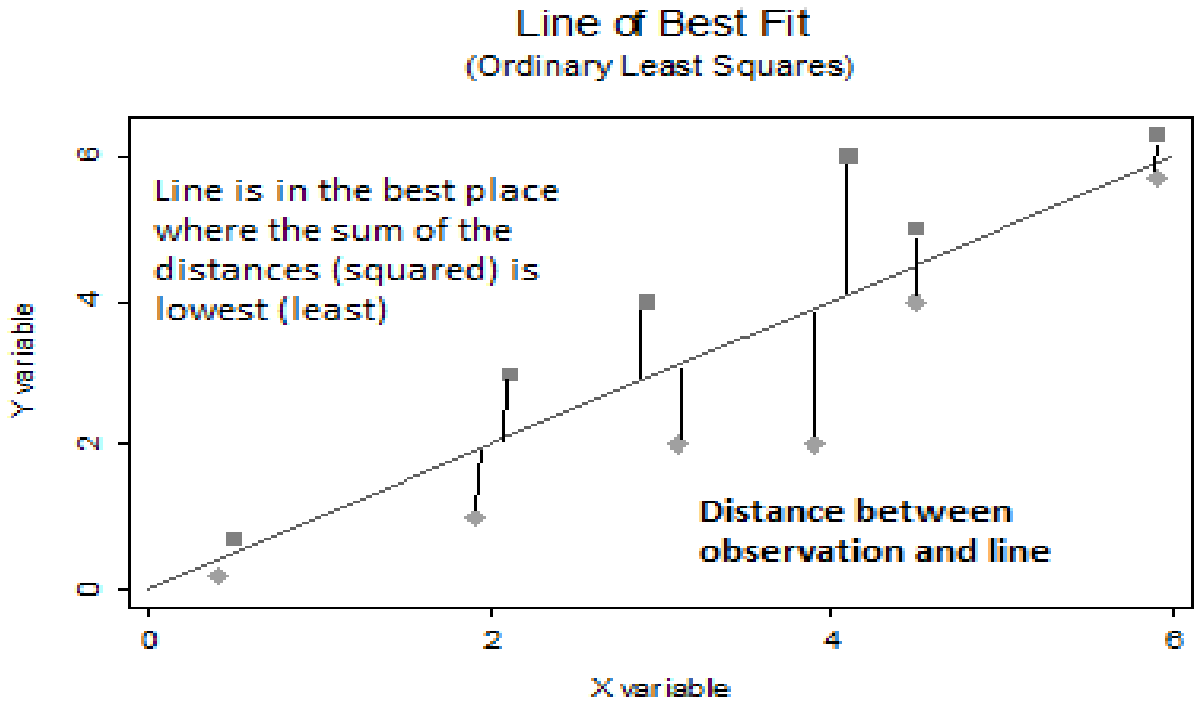
The Average Effect of the Explanatory Variable on the Outcome (β_1)



$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

The Line of Best Fit (Ordinary Least Squares)



Regression Models (elements)

Regression Models (elements)

1. Standard error (s.e.) of β
2. *t* statistic
3. p value
4. Confidence Interval

Standard Error (s.e.)

A measure of the precision with which the regression coefficient is measured (big s.e. indicates imprecision)

(if a coefficient is large compared with its standard error, then it is probably different from 0)

The *t* statistic

$$t = \beta / se$$

Simple guide: when *t* is greater than plus or minus 2 then the variable is significant

And… if β is twice the s.e. the variable is significant

p value

$$t = \beta / se$$

With associated degrees of freedom ($n - k$)

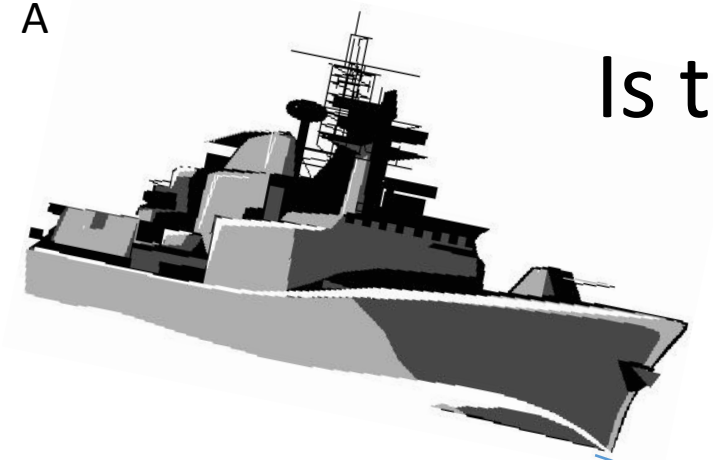
Confidence Intervals

Confidence Intervals (means)

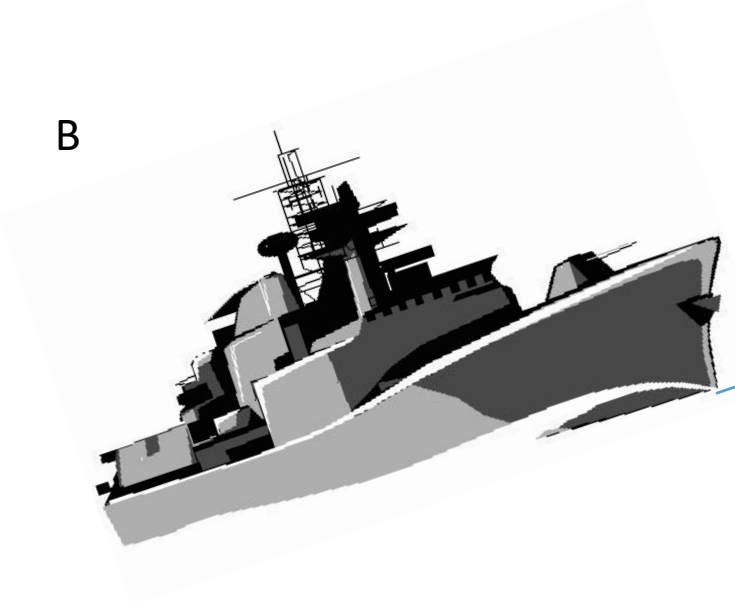
$$CI = \overline{x} \pm (1.96 se_m)$$

Is there a risk of a collision at point c?

A

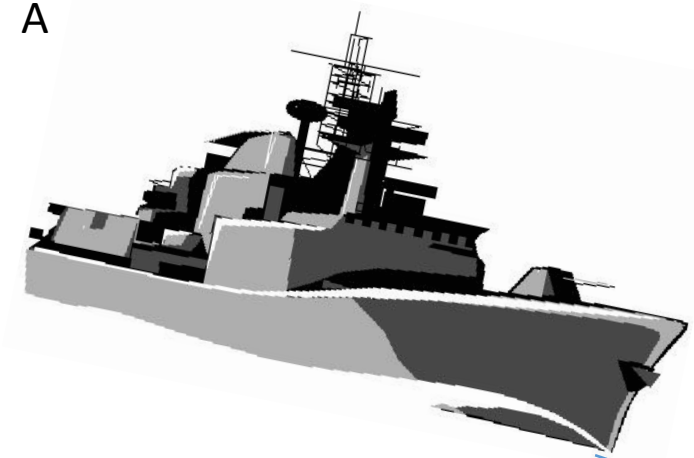


B



C

A

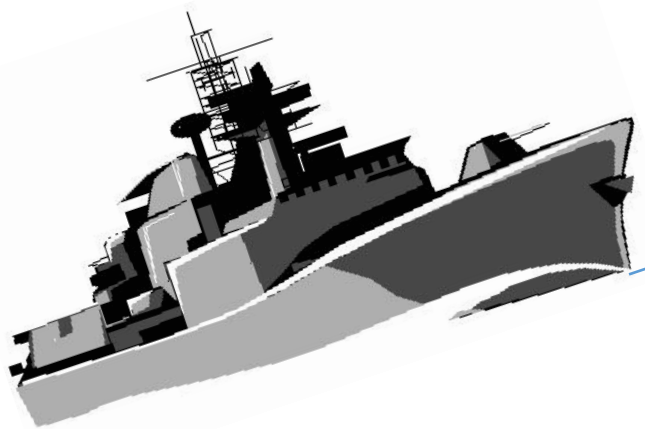


Ship A plans to be at point C at 10:00 am

95% of the time she will arrive between
9:55 am and 10:05 am

C

B

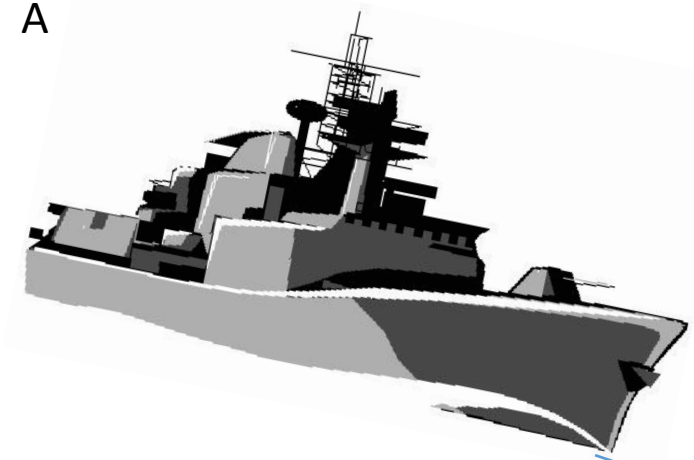


Ship B plans to be at point C at 10:15 am

95% of the time she will arrive between
10:10 am and 10:20 am

On Another Day…

A

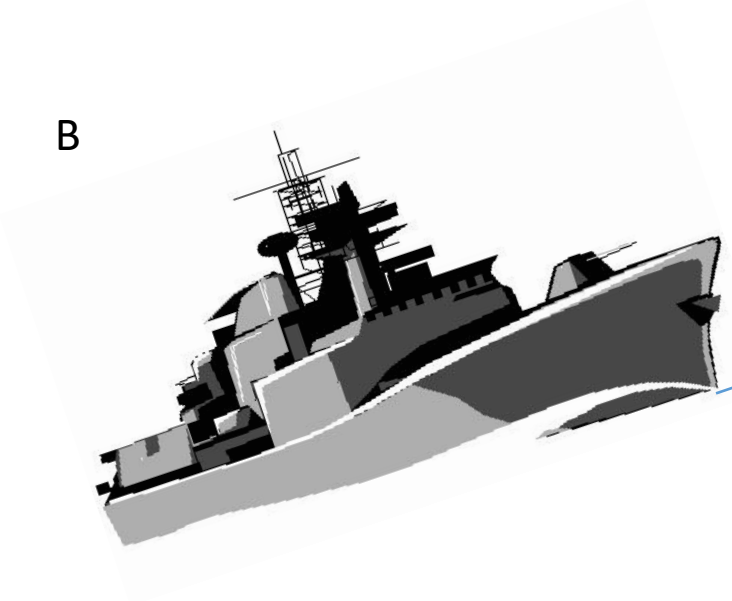


Ship A plans to be at point C at 10:00 am

95% of the time she will arrive between 9:50 am and 10:10 am

C

B



Ship B plans to be at point C at 10:15 am

95% of the time she will arrive between 10:05 am and 10:25am

Confidence Intervals

When confidence intervals for two group statistics (e.g. Mean A and Mean B) overlap then the groups are not significantly different

When there is ‘clear blue water’ there is a significant difference

Confidence Intervals β

$$CI = \beta \pm (1.96 se_{\beta})$$

Confidence Intervals

When confidence interval for β overlaps (i.e. includes) 0 then the coefficient is not significant

There has to be ‘clear blue water’ between β and 0

Regression Models (outputs)

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	=	200	
F(2, 197)	=	68.38	
Prob > F	=	0.0000	
R-squared	=	0.4098	
Adj R-squared	=	0.4038	
Root MSE	=	7.645	

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	=	200	
F(2, 197)	=	68.38	
Prob > F	=	0.0000	
R-squared	=	0.4098	
Adj R-squared	=	0.4038	
Root MSE	=	7.645	

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	=	200
F(2, 197)	=	68.38
Prob > F	=	0.0000
R-squared	=	0.4098
Adj R-squared	=	0.4038
Root MSE	=	7.645

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	=	200	
F(2, 197)	=	68.38	
Prob > F	=	0.0000	
R-squared	=	0.4098	
Adj R-squared	=	0.4038	
Root MSE	=	7.645	

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

Source	SS	df	MS		
Model	7993.54995	2	3996.77498	Prob > F	0.0000
Residual	11513.95	197	58.4464469	R-squared	0.4098
				Adj R-squared	0.4038
Total	19507.5	199	98.0276382	Root MSE	7.645

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	=	200	
F(2, 197)	=	68.38	
Prob > F	=	0.0000	
R-squared	=	0.4098	
Adj R-squared	=	0.4038	
Root MSE	=	7.645	

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	=	200	
F(2, 197)	=	68.38	
Prob > F	=	0.0000	
R-squared	=	0.4098	
Adj R-squared	=	0.4038	
Root MSE	=	7.645	

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	=	200	
F(2, 197)	=	68.38	
Prob > F	=	0.0000	
R-squared	=	0.4098	
Adj R-squared	=	0.4038	
Root MSE	=	7.645	

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

The *t* statistic

$$t = \beta / se$$

$$t = -2.17 / 1.09$$

$$t = -2.00$$

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	=	200	
F(2, 197)	=	68.38	
Prob > F	=	0.0000	
R-squared	=	0.4098	
Adj R-squared	=	0.4038	
Root MSE	=	7.645	

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

Confidence Intervals β

$$CI = \beta \pm (1.96 se_{\beta})$$

$$CI = -2.17 \pm (1.96 * 1.09)$$

Confidence Intervals β

$$CI = \beta \pm (1.96 se_{\beta})$$

$$CI = -2.17 \pm (1.96 * 1.09)$$

$$[-4.31, \quad -0.03]$$

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	=	200	
F(2, 197)	=	68.38	
Prob > F	=	0.0000	
R-squared	=	0.4098	
Adj R-squared	=	0.4038	
Root MSE	=	7.645	

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

When confidence interval for β overlaps 0 (i.e. includes 0)
then the coefficient is not significant

There has to be 'clear blue water' between β and 0

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	=	200	
F(2, 197)	=	68.38	
Prob > F	=	0.0000	
R-squared	=	0.4098	
Adj R-squared	=	0.4038	
Root MSE	=	7.645	

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

Goodness of Fit

```
      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(6, 193)       =      8.17
       Model |  3621.08648         6  603.514413   Prob > F        =    0.0000
    Residual |  14257.7885       193  73.8745519   R-squared       =    0.2025
-------------+----------------------------------   Adj R-squared   =    0.1777
       Total |   17878.875       199   89.843593   Root MSE        =     8.595

------------------------------------------------------------------------------
       write | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |   4.993606   1.243548     4.02   0.000     2.540917    7.446295
    hispanic |  -6.466839   1.929103    -3.35   0.001    -10.27167   -2.662008
       asian |   3.263157   2.701947     1.21   0.229    -2.065978    8.592292
     africam |  -5.482071   2.131616    -2.57   0.011    -9.686324   -1.277818
 middleclass |  -3.651237   1.435617    -2.54   0.012     -6.48275   -.8197234
  lowerclass |  -4.447549   1.772118    -2.51   0.013    -7.942753   -.9523443
       _cons |   53.97775   1.309698    41.21   0.000     51.39459    56.56091
------------------------------------------------------------------------------
```

Source	SS	df	MS	Number of obs	=	200
				F(6, 193)	=	8.17
Model	3621.08648	6	603.514413	Prob > F	=	0.0000
Residual	14257.7885	193	73.8745519	R-squared	=	0.2025
				Adj R-squared	=	0.1777
Total	17878.875	199	89.843593	Root MSE	=	8.595

```
      Source |       SS             df       MS      Number of obs   =       200
-------------+------------------------------------   F(6, 193)       =      8.17
       Model |  3621.08648           6  603.514413   Prob > F        =    0.0000
    Residual |  14257.7885         193  73.8745519   R-squared       =    0.2025
-------------+------------------------------------   Adj R-squared   =    0.1777
       Total |   17878.875         199   89.843593   Root MSE        =     8.595
```

R^2 = Model SS / Total SS

3621.08648 / 17878.875 = .20253436

```
      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(6, 193)       =      8.17
       Model |  3621.08648         6  603.514413   Prob > F        =    0.0000
    Residual |  14257.7885       193  73.8745519   R-squared       =    0.2025
-------------+----------------------------------   Adj R-squared   =    0.1777
       Total |   17878.875       199   89.843593   Root MSE        =     8.595
```

Adjusted R Squared $1-((1 - R^2)(n-1) / (n- k – 1))$

Anova table

Source	SS	df	MS		
Model	3621.08648	6	603.514413	Prob > F	0.0000
Residual	14257.7885	193	73.8745519	R-squared	0.2025
Total	17878.875	199	89.843593	Root MSE	8.595

Number of obs = 200
F(6, 193) = 8.17
Adj R-squared = 0.1777

```
      Source |       SS             df       MS      Number of obs   =       200
-------------+------------------------------------   F(6, 193)       =      8.17
       Model |  3621.08648           6  603.514413   Prob > F        =    0.0000
    Residual |  14257.7885         193  73.8745519   R-squared       =    0.2025
-------------+------------------------------------   Adj R-squared   =    0.1777
       Total |  17878.875          199   89.843593   Root MSE        =     8.595
```

```
      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(6, 193)       =      8.17
       Model |  3621.08648         6  603.514413   Prob > F        =    0.0000
    Residual |  14257.7885       193  73.8745519   R-squared       =    0.2025
-------------+----------------------------------   Adj R-squared   =    0.1777
       Total |   17878.875       199   89.843593   Root MSE        =     8.595
```

```
      Source |       SS             df       MS      Number of obs   =       200
-------------+------------------------------------   F(6, 193)       =      8.17
       Model |  3621.00648           6  603.514413   Prob > F        =    0.0000
    Residual |  14257.7885         193  73.8745519   R-squared       =    0.2025
-------------+------------------------------------   Adj R-squared   =    0.1777
       Total |   17878.875         199   89.843593   Root MSE        =     8.595
```

Source	SS	df	MS		
Model	3621.08648	6	**603.514413**	Prob > F	0.0000
Residual	14257.7885	193	73.8745519	R-squared	0.2025
Total	17878.875	199	89.843593	Root MSE	8.595

Number of obs	200
F(6, 193)	8.17
Adj R-squared	0.1777

```
      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(6, 193)       =      8.17
       Model |  3621.08648         6  603.514413   Prob > F        =    0.0000
    Residual |  14257.7885       193  73.8745519   R-squared       =    0.2025
-------------+----------------------------------   Adj R-squared   =    0.1777
       Total |   17878.875       199  89.843593    Root MSE        =     8.595
```

Source	SS	df	MS		
Model	3621.08648	6	603.514413	Number of obs	200
Residual	14257.7885	193	73.8745519	F(6, 193)	8.17
Total	17878.875	199	89.843593	Prob > F	0.0000
				R-squared	0.2025
				Adj R-squared	0.1777
				Root MSE	8.595

$$F = \frac{\text{(MS MODEL)}}{\text{(MS RESIDUAL}}$$

F = 603.514413 / 73.8745519

```
      Source |       SS              df       MS      Number of obs   =       200
-------------+------------------------------          F(6, 193)       =      8.17
       Model |  3621.08648             6  603.514413   Prob > F        =    0.0000
    Residual |  14257.7885           193  73.8745519   R-squared       =    0.2025
-------------+------------------------------          Adj R-squared   =    0.1777
       Total |   17878.875           199   89.843593   Root MSE        =     8.595
```

F = 8.17

The Critical Value of F with (6, 193) df = 2.15

Source	SS	df	MS		
Model	3621.08648	6	603.514413	Number of obs	= 200
Residual	14257.7885	193	73.8745519	F(6, 193)	= 8.17
				Prob > F	= 0.0000
				R-squared	= 0.2025
				Adj R-squared	= 0.1777
Total	17878.875	199	89.843593	Root MSE	= 8.595

Square Root (MS Residual)

$$\sqrt{(73.8745519)}$$

Bayesian Information Criterion (BIC)

A measure of model parsimony

1. Comparing non-nested models
2. Correction for large samples
3. Penalizes large models

Bayesian Information Criterion (BIC)

In general,

statistical models with a lower BIC value are preferable because they are more parsimonious descriptions of the data

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

Modelling Binary Outcomes

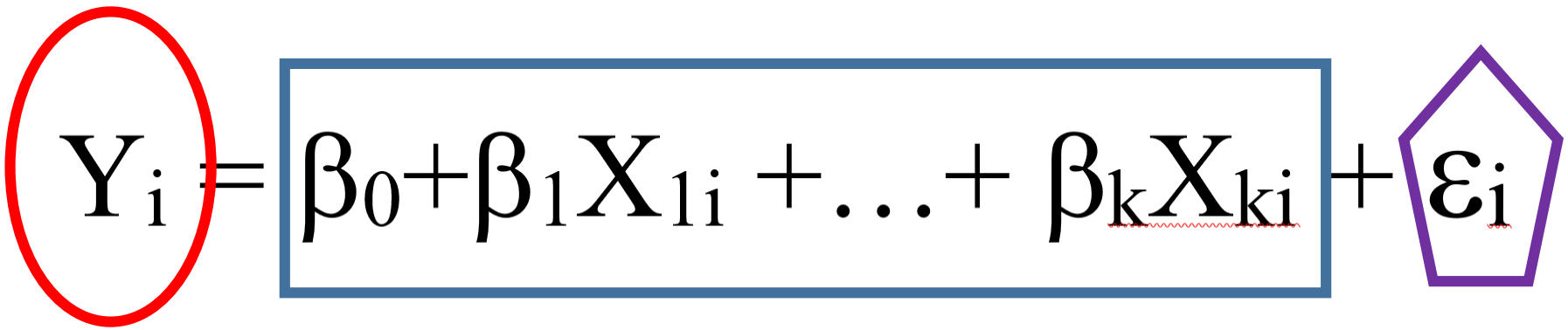
Photo: https://www.anp-archief.nl/page/133774/leeuw-edith-de

Photo: https://www.yuppiechef.com/spatula/myth-or-magic-does-toast-always-land-butter-side-down/

Binary Outcome (0,1)

Examples are legion

- Yes / No
- Present / Absent
- Pass / Fail
- Pregnant / Not Pregnant


$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$

Modelling Binary Outcome (0,1)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Y is either 0 or 1

Chance of 1 rather than 0

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Y is either 0 or 1

Chance of 1 rather than 0

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Alternative error structure

Model Components

R^2

β

Standard error (s.e.) of β

t statistic

p value

Confidence Interval

First Year Test (Pass = 1; Fail = 0)

- n = 189
- Hours spent on gaming consul during semester
- Gender (male / female)

```
. logit pass hours female

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.72132
Iteration 2:   log likelihood = -110.58286
Iteration 3:   log likelihood = -110.58272
Iteration 4:   log likelihood = -110.58272

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =      13.51
                                                Prob > chi2       =     0.0012
Log likelihood = -110.58272                     Pseudo R2         =     0.0576

------------------------------------------------------------------------------
        pass |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0186285   .0065928    -2.83   0.005    -.0315502   -.0057068
      female |   1.854477   .7008247     2.65   0.008     .4808862    3.228068
       _cons |   1.447862   .8208979     1.76   0.078    -.1610682    3.056792
------------------------------------------------------------------------------
```

```
. logit pass hours female

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.72132
Iteration 2:   log likelihood = -110.58286
Iteration 3:   log likelihood = -110.58272
Iteration 4:   log likelihood = -110.58272

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =      13.51
                                                Prob > chi2       =     0.0012
Log likelihood = -110.58272                     Pseudo R2         =     0.0576
```

pass	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hours	-.0186285	.0065928	-2.83	0.005	-.0315502	-.0057068
female	1.854477	.7008247	2.65	0.008	.4808862	3.228068
_cons	1.447862	.8208979	1.76	0.078	-.1610682	3.056792

```
. logit pass hours female

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.72132
Iteration 2:   log likelihood = -110.58286
Iteration 3:   log likelihood = -110.58272
Iteration 4:   log likelihood = -110.58272

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =      13.51
                                                Prob > chi2       =     0.0012
Log likelihood = -110.58272                     Pseudo R2         =     0.0576
```

pass	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hours	-.0186285	.0065928	-2.83	0.005	-.0315502	-.0057068
female	1.854477	.7008247	2.65	0.008	.4808862	3.228068
_cons	1.447862	.8208979	1.76	0.078	-.1610682	3.056792

```
. logit pass hours female

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.72132
Iteration 2:   log likelihood = -110.58286
Iteration 3:   log likelihood = -110.58272
Iteration 4:   log likelihood = -110.58272

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =      13.51
                                                Prob > chi2       =     0.0012
Log likelihood = -110.58272                     Pseudo R2         =     0.0576

------------------------------------------------------------------------------
        pass |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0186285   .0065928    -2.83   0.005    -.0315502   -.0057068
      female |   1.854477   .7008247     2.65   0.008     .4808862    3.228068
       _cons |   1.447862   .8208979     1.76   0.078    -.1610682    3.056792
------------------------------------------------------------------------------
```

```
. logit pass hours female

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.72132
Iteration 2:   log likelihood = -110.58286
Iteration 3:   log likelihood = -110.58272
Iteration 4:   log likelihood = -110.58272

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =      13.51
                                                Prob > chi2       =     0.0012
Log likelihood = -110.58272                     Pseudo R2         =     0.0576
```

pass	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hours	-.0186285	.0065928	-2.83	0.005	-.0315502	-.0057068
female	1.854477	.7008247	2.65	0.008	.4808862	3.228068
_cons	1.447862	.8208979	1.76	0.078	-.1610682	3.056792

Probabilities

Probability of rolling a six

There are six sides

1 / 6 = .17 or 17%

Odds

Odds of rolling a six

There is 1 side with a 6
and 5 sides without a 6

1 : 5 = .20

Probabilities & Odds

Probability = odds / (1 + odds)

.2 / (1 + .2) = .17

Odds & Probabilities

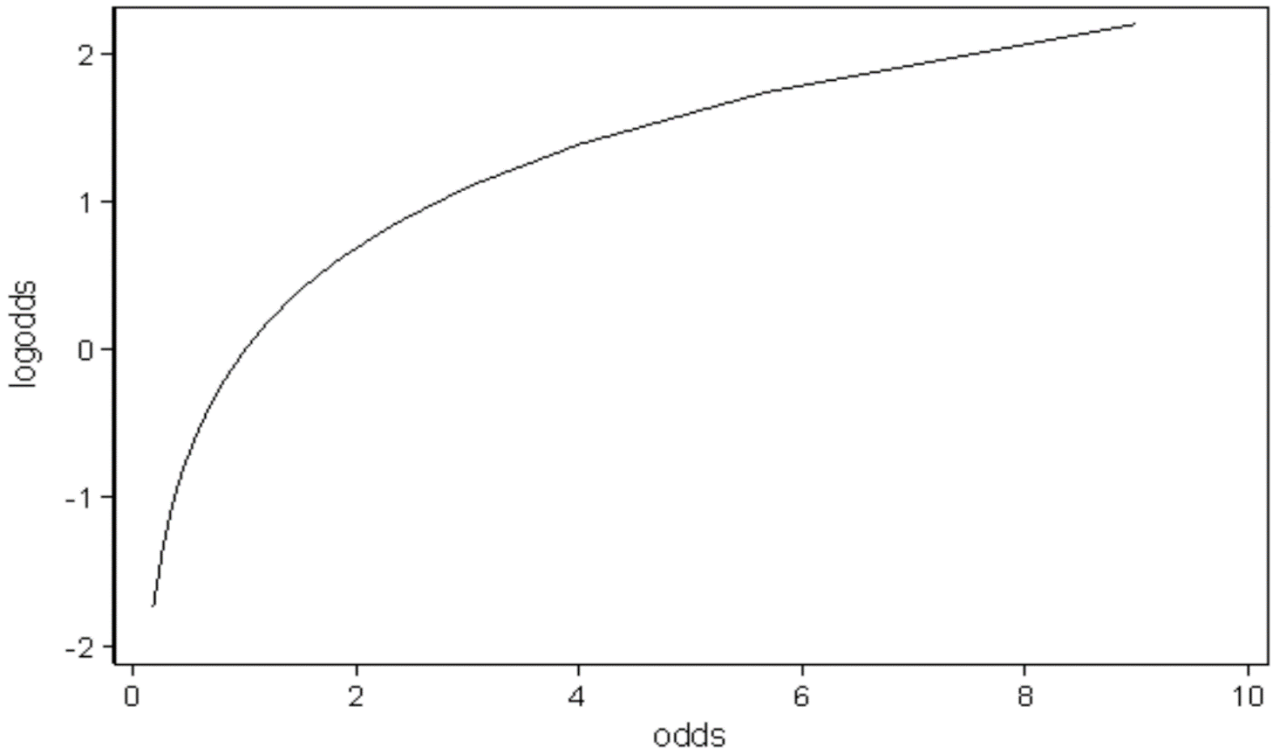
Odds = p / (1 - p)

.17 / (1 - .17) = .20

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

β estimated on the log odds scale

The logistic regression model uses a transformation from odds to log of odds



$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Estimates the effect a one unit change in X_1 has on the log odds of Y=1

Standard Error (s.e.)

A measure of the precision with which the regression coefficient is measured (big s.e. indicates imprecision)

(if a coefficient is large compared with its standard error, then it is probably different from 0)

Standard Error (s.e.)

A measure of the precision with which the regression coefficient is measured (big s.e. indicates imprecision)

(if a coefficient is large compared with its standard error, then it is probably different from 0)

Beware in a logistic regression model s.e. will also be on the log odds scale

```
. logit pass hours female

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.72132
Iteration 2:   log likelihood = -110.58286
Iteration 3:   log likelihood = -110.58272
Iteration 4:   log likelihood = -110.58272

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =      13.51
                                                Prob > chi2       =     0.0012
Log likelihood = -110.58272                     Pseudo R2         =     0.0576

------------------------------------------------------------------------------
        pass |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0186285   .0065928    -2.83   0.005    -.0315502   -.0057068
      female |   1.854477   .7008247     2.65   0.008     .4808862    3.228068
       _cons |   1.447862   .8208979     1.76   0.078    -.1610682    3.056792
------------------------------------------------------------------------------
```

Z Statistic

$$z = \beta / se$$

Simple rule: when *z* is greater than plus or minus 2 then the variable is significant

And… if β is twice the s.e. the variable is significant

```
. logit pass hours female

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.72132
Iteration 2:   log likelihood = -110.58286
Iteration 3:   log likelihood = -110.58272
Iteration 4:   log likelihood = -110.58272

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =      13.51
                                                Prob > chi2       =     0.0012
Log likelihood = -110.58272                     Pseudo R2         =     0.0576
```

pass	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hours	-.0186285	.0065928	-2.83	0.005	-.0315502	-.0057068
female	1.854477	.7008247	2.65	0.008	.4808862	3.228068
_cons	1.447862	.8208979	1.76	0.078	-.1610682	3.056792

p value

$$z = \beta / se$$

```
. logit pass hours female

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.72132
Iteration 2:   log likelihood = -110.58286
Iteration 3:   log likelihood = -110.58272
Iteration 4:   log likelihood = -110.58272

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =      13.51
                                                Prob > chi2       =     0.0012
Log likelihood = -110.58272                     Pseudo R2         =     0.0576
```

pass	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hours	-.0186285	.0065928	-2.83	0.005	-.0315502	-.0057068
female	1.854477	.7008247	2.65	0.008	.4808862	3.228068
_cons	1.447862	.8208979	1.76	0.078	-.1610682	3.056792

Confidence Intervals β

$$CI = \beta \pm (1.96 se_{\beta})$$

```
. logit pass hours female

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.72132
Iteration 2:   log likelihood = -110.58286
Iteration 3:   log likelihood = -110.58272
Iteration 4:   log likelihood = -110.58272

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =      13.51
                                                Prob > chi2       =     0.0012
Log likelihood = -110.58272                     Pseudo R2         =     0.0576

------------------------------------------------------------------------------
        pass |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0186285   .0065928    -2.83   0.005    -.0315502   -.0057068
      female |   1.854477   .7008247     2.65   0.008     .4808862    3.228068
       _cons |   1.447862   .8208979     1.76   0.078    -.1610682    3.056792
------------------------------------------------------------------------------
```

Confidence Intervals

When confidence interval for β overlaps (i.e. includes) 0 then the coefficient is not significant

There has to be ‘clear blue water’ between β and 0

```
. logit pass hours female

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.72132
Iteration 2:   log likelihood = -110.58286
Iteration 3:   log likelihood = -110.58272
Iteration 4:   log likelihood = -110.58272

Logistic regression                             Number of obs     =        189
                                                LR chi2(2)        =      13.51
                                                Prob > chi2       =     0.0012
Log likelihood = -110.58272                     Pseudo R2         =     0.0576
```

pass	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hours	-.0186285	.0065928	-2.83	0.005	-.0315502	-.0057068
female	1.854477	.7008247	2.65	0.008	.4808862	3.228068
_cons	1.447862	.8208979	1.76	0.078	-.1610682	3.056792

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

Interpreting Coefficients

Original Article

Statistical modelling of key variables in social survey data analysis

Methodological Innovations
Volume 9: 1–17

DOI: 10.1177/2059799116638002
mio.sagepub.com
SAGE

Roxanne Connelly[1], Vernon Gayle[2] and Paul S. Lambert[3]

Abstract

The application of statistical modelling techniques has become a cornerstone of analyses of large-scale social survey data. Bringing this special section on key variables to a close, this final article discusses several important issues relating to the inclusion of key variables in statistical modelling analyses. We outline two, often neglected, issues that are relevant to a great many applications of statistical models based upon social survey data. The first is known as the reference category problem and is related to the interpretation of categorical explanatory variables. The second is the interpretation and comparison of the effects from models for non-linear outcomes. We then briefly discuss other common complexities in using statistical models for social science research; these include the non-linear transformation of variables, and considerations of intersectionality and interaction effects. We conclude by emphasising the importance of two, often overlooked, elements of the social survey data analysis process, sensitivity analysis and documentation for replication. We argue that more attention should routinely be devoted to these issues.

Table 1. An example of a linear regression model (model 1) and a logistic regression model (model 2). The outcome of model 1 is score on a maths test. The outcome of model 2 is the score on the same maths test categorised into above average (1) and average and below attainment (0).

		Model 1	Model 2
		Linear regression coefficients (standard errors)	Logistic regression coefficients log odds (standard errors)
Father's NS-SEC	1. Large employers, higher managerial and professional	Ref.	Ref.
	2. Lower managerial and professional	−2.09*** (0.50)	−0.30** (0.10)
	3. Intermediate	−2.52*** (0.55)	−0.34** (0.11)
	4. Small employers and own account workers	−4.38*** (0.52)	−0.67*** (0.10)
	5. Lower supervisory and technical	−5.34*** (0.48)	−0.82*** (0.09)
	6. Semi-routine	−5.77*** (0.50)	−0.89*** (0.10)
	7. Routine	−6.80*** (0.49)	−0.95*** (0.09)
Home owner		3.53*** (0.27)	0.53*** (0.05)
Mother very interested in child's education		5.29*** (0.26)	0.74*** (0.05)
Constant		44.19 (0.46)	0.01 (0.09)
Adjusted R^2		0.15	
McFadden's adjusted R^2			0.07
McKelvey and Zavoina's R^2			0.13
Cragg and Uhler's R^2			0.13
n		8198	8198

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
1970 British Cohort Study, age 10 survey.

Table 1. An example of a linear regression model (model 1) and a logistic regression model (model 2). The outcome of model 1 is score on a maths test. The outcome of model 2 is the score on the same maths test categorised into above average (1) and average and below attainment (0).

		Model 1	Model 2
		Linear regression coefficients (standard errors)	Logistic regression coefficients log odds (standard errors)
Father's NS-SEC	1. Large employers, higher managerial and professional	Ref.	Ref.
	2. Lower managerial and professional	−2.09*** (0.50)	−0.30** (0.10)
	3. Intermediate	−2.52*** (0.55)	−0.34** (0.11)
	4. Small employers and own account workers	−4.38*** (0.52)	−0.67*** (0.10)
	5. Lower supervisory and technical	−5.34*** (0.48)	−0.82*** (0.09)
	6. Semi-routine	−5.77*** (0.50)	−0.89*** (0.10)
	7. Routine	−6.80*** (0.49)	−0.95*** (0.09)
Home owner		3.53*** (0.27)	0.53*** (0.05)
Mother very interested in child's education		5.29*** (0.26)	0.74*** (0.05)
Constant		44.19 (0.46)	0.01 (0.09)
Adjusted R^2		0.15	
McFadden's adjusted R^2			0.07
McKelvey and Zavoina's R^2			0.13
Cragg and Uhler's R^2			0.13
n		8198	8198

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
1970 British Cohort Study, age 10 survey.

Table 1. An example of a linear regression model (model 1) and a logistic regression model (model 2). The outcome of model 1 is score on a maths test. The outcome of model 2 is the score on the same maths test categorised into above average (1) and average and below attainment (0).

		Model 1	Model 2
		Linear regression coefficients (standard errors)	Logistic regression coefficients log odds (standard errors)
Father's NS-SEC	1. Large employers, higher managerial and professional	Ref.	Ref.
	2. Lower managerial and professional	−2.09*** (0.50)	−0.30** (0.10)
	3. Intermediate	−2.52*** (0.55)	−0.34** (0.11)
	4. Small employers and own account workers	−4.38*** (0.52)	−0.67*** (0.10)
	5. Lower supervisory and technical	−5.34*** (0.48)	−0.82*** (0.09)
	6. Semi-routine	−5.77*** (0.50)	−0.89*** (0.10)
	7. Routine	−6.80*** (0.49)	−0.95*** (0.09)
Home owner		3.53*** (0.27)	0.53*** (0.05)
Mother very interested in child's education		5.29*** (0.26)	0.74*** (0.05)
Constant		44.19 (0.46)	0.01 (0.09)
Adjusted R^2		0.15	
McFadden's adjusted R^2			0.07
McKelvey and Zavoina's R^2			0.13
Cragg and Uhler's R^2			0.13
n		8198	8198

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
1970 British Cohort Study, age 10 survey.

Table 1. An example of a linear regression model (model 1) and a logistic regression model (model 2). The outcome of model 1 is score on a maths test. The outcome of model 2 is the score on the same maths test categorised into above average (1) and average and below attainment (0).

		Model 1	Model 2
		Linear regression coefficients (standard errors)	Logistic regression coefficients log odds (standard errors)
Father's NS-SEC	1. Large employers, higher managerial and professional	Ref.	Ref.
	2. Lower managerial and professional	−2.09*** (0.50)	−0.30** (0.10)
	3. Intermediate	−2.52*** (0.55)	−0.34** (0.11)
	4. Small employers and own account workers	−4.38*** (0.52)	−0.67*** (0.10)
	5. Lower supervisory and technical	−5.34*** (0.48)	−0.82*** (0.09)
	6. Semi-routine	−5.77*** (0.50)	−0.89*** (0.10)
	7. Routine	−6.80*** (0.49)	−0.95*** (0.09)
Home owner		3.53*** (0.27)	0.53*** (0.05)
Mother very interested in child's education		5.29*** (0.26)	0.74*** (0.05)
Constant		44.19 (0.46)	0.01 (0.09)
Adjusted R^2		0.15	
McFadden's adjusted R^2			0.07
McKelvey and Zavoina's R^2			0.13
Cragg and Uhler's R^2			0.13
n		8198	8198

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
1970 British Cohort Study, age 10 survey.

Table 1. An example of a linear regression model (model 1) and a logistic regression model (model 2). The outcome of model 1 is score on a maths test. The outcome of model 2 is the score on the same maths test categorised into above average (1) and average and below attainment (0).

		Model 1	Model 2
		Linear regression coefficients (standard errors)	Logistic regression coefficients log odds (standard errors)
Father's NS-SEC	1. Large employers, higher managerial and professional	Ref.	Ref.
	2. Lower managerial and professional	−2.09*** (0.50)	−0.30** (0.10)
	3. Intermediate	−2.52*** (0.55)	−0.34** (0.11)
	4. Small employers and own account workers	−4.38*** (0.52)	−0.67*** (0.10)
	5. Lower supervisory and technical	−5.34*** (0.48)	−0.82*** (0.09)
	6. Semi-routine	−5.77*** (0.50)	−0.89*** (0.10)
	7. Routine	−6.80*** (0.49)	−0.95*** (0.09)
Home owner		3.53*** (0.27)	0.53*** (0.05)
Mother very interested in child's education		5.29*** (0.26)	0.74*** (0.05)
Constant		44.19 (0.46)	0.01 (0.09)
Adjusted R^2		0.15	
McFadden's adjusted R^2			0.07
McKelvey and Zavoina's R^2			0.13
Cragg and Uhler's R^2			0.13
n		8198	8198

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
1970 British Cohort Study, age 10 survey.

Table 2. An example of the different presentation of parameter estimates from a logistic regression model of maths test scores (model 2 in Table 1).

Father's NS-SEC	1	2
	Log odds	Odds ratio
1. Large employers, higher managerial and professional	Ref.	Ref.
2. Lower managerial and professional	−0.30**	0.74**
3. Intermediate	−0.34**	0.71**
4. Small employers and own account workers	−0.67***	0.51***
5. Lower supervisory and technical	−0.82***	0.44***
6. Semi-routine	−0.89***	0.41***
7. Routine	−0.95***	0.39***

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 3. Conversion of log odds, odds and probabilities.

1	2	3
Odds	Log odds (logit scale)	Probabilities
99.00	4.60	0.99
19.00	2.94	0.95
9.00	2.20	0.90
4.00	1.39	0.80
2.33	0.85	0.70
1.50	0.41	0.60
1.00	0.00	0.50
0.67	−0.41	0.40
0.43	−0.85	0.30
0.25	−1.39	0.20
0.11	−2.20	0.10
0.05	−2.94	0.05
0.01	−4.60	0.01

Table 3. Conversion of log odds, odds and probabilities.

1	2	3
Odds	Log odds (logit scale)	Probabilities
99.00	4.60	0.99
19.00	2.94	0.95
9.00	2.20	0.90
4.00	1.39	0.80
2.33	0.85	0.70
1.50	0.41	0.60
1.00	0.00	0.50
0.67	−0.41	0.40
0.43	−0.85	0.30
0.25	−1.39	0.20
0.11	−2.20	0.10
0.05	−2.94	0.05
0.01	−4.60	0.01

Table 3. Conversion of log odds, odds and probabilities.

1	2	3
Odds	Log odds (logit scale)	Probabilities
99.00	4.60	0.99
19.00	2.94	0.95
9.00	2.20	0.90
4.00	1.39	0.80
2.33	0.85	0.70
1.50	0.41	0.60
1.00	0.00	0.50
0.67	−0.41	0.40
0.43	−0.85	0.30
0.25	−1.39	0.20
0.11	−2.20	0.10
0.05	−2.94	0.05
0.01	−4.60	0.01

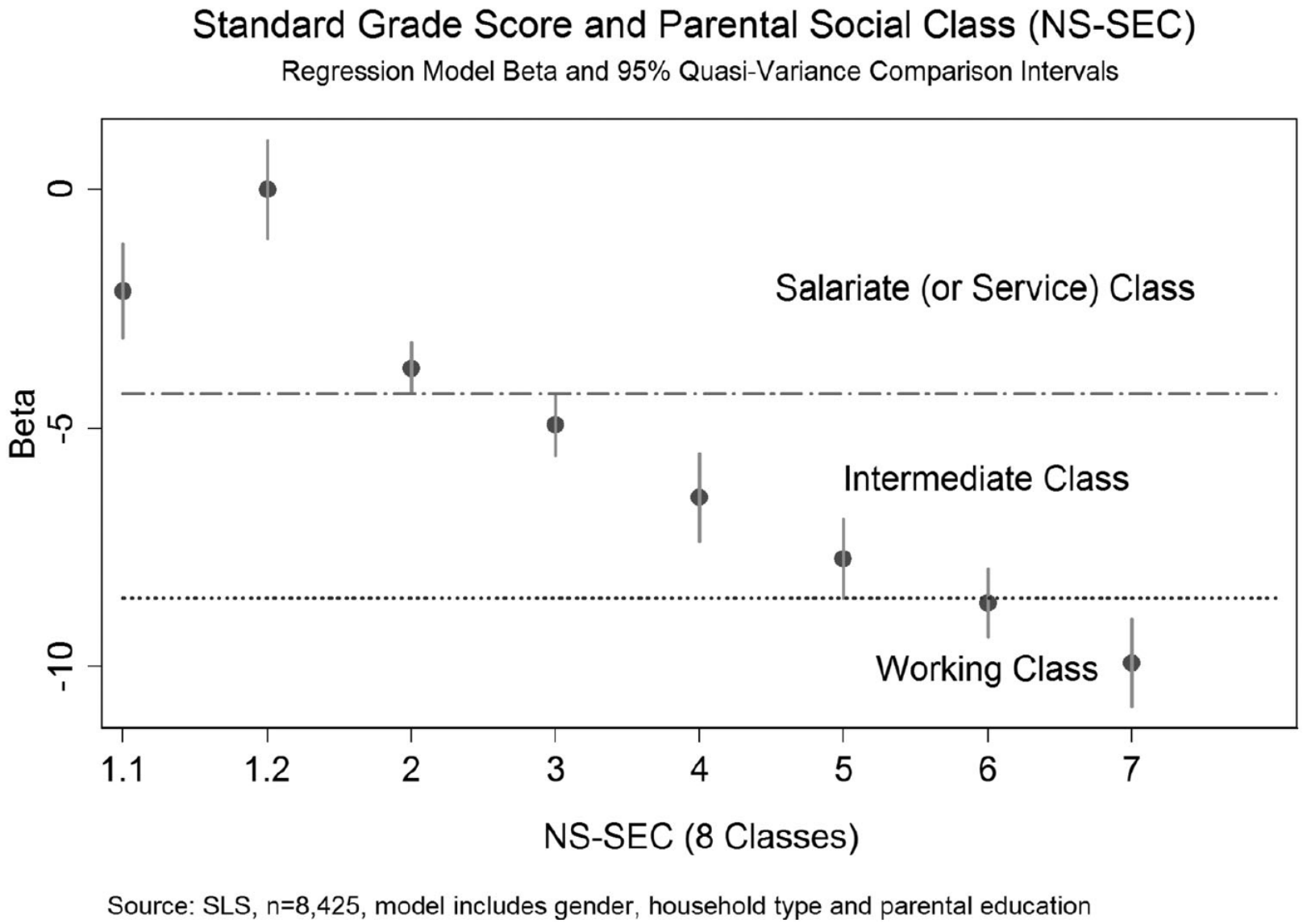


Figure 1. School standard grade scores – parental social class effects (NS-SEC) Model 3.

Table 2. An example of the different presentation of parameter estimates from a logistic regression model of maths test scores (model 2 in Table 1).

Father's NS-SEC	1	2	3	4	5
	Log odds	Odds ratio	Gelman and Hill (probability)	Conditional marginal effects (probability)	Adjusted prediction (probability)
1. Large employers, higher managerial and professional	Ref.	Ref.	Ref.	Ref.	0.67
2. Lower managerial and professional	−0.30**	0.74**	−0.08**	−0.07**	0.60
3. Intermediate	−0.34**	0.71**	−0.09**	−0.08**	0.59
4. Small employers and own account workers	−0.67***	0.51***	−0.17***	−0.16***	0.51
5. Lower supervisory and technical	−0.82***	0.44***	−0.21***	−0.20***	0.47
6. Semi-routine	−0.89***	0.41***	−0.22***	−0.21***	0.46
7. Routine	−0.95***	0.39***	−0.24***	−0.23***	0.44

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 2. An example of the different presentation of parameter estimates from a logistic regression model of maths test scores (model 2 in Table 1).

Father's NS-SEC	1	2	3	4	5
	Log odds	Odds ratio	Gelman and Hill (probability)	Conditional marginal effects (probability)	Adjusted prediction (probability)
1. Large employers, higher managerial and professional	Ref.	Ref.	Ref.	Ref.	0.67
2. Lower managerial and professional	−0.30**	0.74**	−0.08**	−0.07**	0.60
3. Intermediate	−0.34**	0.71**	−0.09**	−0.08**	0.59
4. Small employers and own account workers	−0.67***	0.51***	−0.17***	−0.16***	0.51
5. Lower supervisory and technical	−0.82***	0.44***	−0.21***	−0.20***	0.47
6. Semi-routine	−0.89***	0.41***	−0.22***	−0.21***	0.46
7. Routine	−0.95***	0.39***	−0.24***	−0.23***	0.44

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 2. An example of the different presentation of parameter estimates from a logistic regression model of maths test scores (model 2 in Table 1).

Father's NS-SEC	1	2	3	4	5
	Log odds	Odds ratio	Gelman and Hill (probability)	Conditional marginal effects (probability)	Adjusted prediction (probability)
1. Large employers, higher managerial and professional	Ref.	Ref.	Ref.	Ref.	0.67
2. Lower managerial and professional	−0.30**	0.74**	−0.08**	−0.07**	0.60
3. Intermediate	−0.34**	0.71**	−0.09**	−0.08**	0.59
4. Small employers and own account workers	−0.67***	0.51***	−0.17***	−0.16***	0.51
5. Lower supervisory and technical	−0.82***	0.44***	−0.21***	−0.20***	0.47
6. Semi-routine	−0.89***	0.41***	−0.22***	−0.21***	0.46
7. Routine	−0.95***	0.39***	−0.24***	−0.23***	0.44

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Gelman and Hill (2008) suggest that as a rule of convenience, analysts should take logistic regression coefficients (other than the constant term) and divide them by 4 to get an upper bound of the predictive difference corresponding to a one unit change in the explanatory variable.

Table 2. An example of the different presentation of parameter estimates from a logistic regression model of maths test scores (model 2 in Table 1).

Father's NS-SEC	1	2	3	4	5
	Log odds	Odds ratio	Gelman and Hill (probability)	Conditional marginal effects (probability)	Adjusted prediction (probability)
1. Large employers, higher managerial and professional	Ref.	Ref.	Ref.	Ref.	0.67
2. Lower managerial and professional	−0.30**	0.74**	−0.08**	−0.07**	0.60
3. Intermediate	−0.34**	0.71**	−0.09**	−0.08**	0.59
4. Small employers and own account workers	−0.67***	0.51***	−0.17***	−0.16***	0.51
5. Lower supervisory and technical	−0.82***	0.44***	−0.21***	−0.20***	0.47
6. Semi-routine	−0.89***	0.41***	−0.22***	−0.21***	0.46
7. Routine	−0.95***	0.39***	−0.24***	−0.23***	0.44

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Gelman and Hill (2008) suggest that as a rule of convenience, analysts should take logistic regression coefficients (other than the constant term) and divide them by 4 to get an upper bound of the predictive difference corresponding to a one unit change in the explanatory variable.

Table 2. An example of the different presentation of parameter estimates from a logistic regression model of maths test scores (model 2 in Table 1).

Father's NS-SEC	1	2	3	4	5
	Log odds	Odds ratio	Gelman and Hill (probability)	Conditional marginal effects (probability)	Adjusted prediction (probability)
1. Large employers, higher managerial and professional	Ref.	Ref.	Ref.	Ref.	0.67
2. Lower managerial and professional	−0.30**	0.74**	−0.08**	−0.07**	0.60
3. Intermediate	−0.34**	0.71**	−0.09**	−0.08**	0.59
4. Small employers and own account workers	−0.67***	0.51***	−0.17***	−0.16***	0.51
5. Lower supervisory and technical	−0.82***	0.44***	−0.21***	−0.20***	0.47
6. Semi-routine	−0.89***	0.41***	−0.22***	−0.21***	0.46
7. Routine	−0.95***	0.39***	−0.24***	−0.23***	0.44

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 2. An example of the different presentation of parameter estimates from a logistic regression model of maths test scores (model 2 in Table 1).

Father's NS-SEC	1	2	3	4	5
	Log odds	Odds ratio	Gelman and Hill (probability)	Conditional marginal effects (probability)	Adjusted prediction (probability)
1. Large employers, higher managerial and professional	Ref.	Ref.	Ref.	Ref.	0.67
2. Lower managerial and professional	−0.30**	0.74**	−0.08**	−0.07**	0.60
3. Intermediate	−0.34**	0.71**	−0.09**	−0.08**	0.59
4. Small employers and own account workers	−0.67***	0.51***	−0.17***	−0.16***	0.51
5. Lower supervisory and technical	−0.82***	0.44***	−0.21***	−0.20***	0.47
6. Semi-routine	−0.89***	0.41***	−0.22***	−0.21***	0.46
7. Routine	−0.95***	0.39***	−0.24***	−0.23***	0.44

NS-SEC: National Statistics Socio-Economic Classification. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

We have found this simple technique to be especially useful when we are in the audience at seminars and conference presentations.

Model Components (linear regression)

1. R^2
2. β
3. Standard error (s.e.) of β
4. *t* statistic
5. p value
6. Confidence Interval

Model Components (logistic regression)

1. R^2 pseudo
2. β (log odds)
3. Standard error (s.e.) of β
4. z statistic
5. p value
6. Confidence Interval
7. Odd ratios
8. Marginal effects

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

The Generalize Linear Modelling Family (glm)

Multinomial Logit Models

Model Components

1. R^2
2. β
3. Standard error (s.e.) of β
4. A *t* or *z* statistic
5. p value
6. Confidence Interval

mlogit Model

Categorical Outcome (0,1,2,3…)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Y is either 0,1,2,3

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Estimates the effect a one unit change in X_1 on the log odds scale (like logit)

Mlogit Model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Alternative error structure (generalized logit)

Model Components

1. R^2
2. β (log odds)
3. Standard error (s.e.) of β
4. A *t* or *z* statistic
5. p value
6. Confidence Interval

Table 5. Multinomial Logistic regression on GCSE (A*–C) Attainment Categories.

	None (No GCSEs A*-C) *versus* **Middle** (1-4 GCSEs A*-C)			**Benchmark** (5+ GCSEs Grade A*-C) *versus* **Middle** (1-4 GCSEs A*-C)		
	Coefficient		**SE**	**Coefficient**		**SE**
Gender						
Female	0.00		(0.00)	0.00		(0.00)
Male	-0.06		(0.22)	-0.60	**	(0.21)
Parental Social Class						
Routine/Manual	0.00		(0.00)	0.00		(0.00)
Intermediate	-0.21		(0.30)	0.02		(0.31)
Managerial/Professional	-0.61	*	(0.27)	-0.03		(0.27)
School Year	-0.05		(0.04)	-0.04		(0.04)
Parental Education						
None	0.00		(0.00)	0.00		(0.00)
Sub-Degree (e.g. school level qualifications)	0.13		(0.33)	0.82	*	(0.37)
Degree Level	0.59		(0.45)	1.66	***	(0.47)
Home Owners						
Renters (Private & Local Authority)	0.00		(0.00)	0.00		(0.00)
Home Owners	-0.41		(0.29)	0.32		(0.32)
Constant	1.46	***	(0.41)	0.23		(0.46)
Log-likelihood	-706.72					
Nagelkerke R^2	0.11					
McFadden's Adjusted R^2	0.01					
Total Number of Observations	713					

Notes: British Household Panel Survey 'rising 16s', England and Wales, unweighted data.

Connelly, R., Murray, S. and Gayle, V. (2013) 'Young People and School Attainment: Exploring the 'Middle'', *Sociological Research Online,* 18(1)6, http://www.socresonline.org.uk/18/1/6.html

Table 5. Multinomial Logistic regression on GCSE (A*–C) Attainment Categories.

	None (No GCSEs A*-C) *versus* Middle (1-4 GCSEs A*-C)			Benchmark (5+ GCSEs Grade A*-C) *versus* Middle (1-4 GCSEs A*-C)		
	Coefficient		**SE**	**Coefficient**		**SE**
Gender						
Female	0.00		(0.00)	0.00		(0.00)
Male	-0.06		(0.22)	-0.60	**	(0.21)
Parental Social Class						
Routine/Manual	0.00		(0.00)	0.00		(0.00)
Intermediate	-0.21		(0.30)	0.02		(0.31)
Managerial/Professional	-0.61	*	(0.27)	-0.03		(0.27)
School Year	-0.05		(0.04)	-0.04		(0.04)
Parental Education						
None	0.00		(0.00)	0.00		(0.00)
Sub-Degree (e.g. school level qualifications)	0.13		(0.33)	0.82	*	(0.37)
Degree Level	0.59		(0.45)	1.66	***	(0.47)
Home Owners						
Renters (Private & Local Authority)	0.00		(0.00)	0.00		(0.00)
Home Owners	-0.41		(0.29)	0.32		(0.32)
Constant	1.46	***	(0.41)	0.23		(0.46)
Log-likelihood	-706.72					
Nagelkerke R^2	0.11					
McFadden's Adjusted R^2	0.01					
Total Number of Observations	713					

Notes: British Household Panel Survey 'rising 16s', England and Wales, unweighted data.

Connelly, R., Murray, S. and Gayle, V. (2013) 'Young People and School Attainment: Exploring the 'Middle'', *Sociological Research Online,* 18(1)6, http://www.socresonline.org.uk/18/1/6.html

Ordinal (Logit) Models

A large amount of data analysed within social science studies consists of categorical outcome variables that can plausibly be considered as having a substantively interesting order (for example levels of attainment in educational qualifications)

Gayle, V., 1996. Modelling tabular data with an ordered outcome. *Sociological Research Online*, *1*(3), pp.1-10.

Ordinal (Logit) Models

Ologit model (proportional odds model)

Continuation ratio model

Model Components

1. R^2
2. β (log odds)
3. Standard error (s.e.) of β
4. A *t* or *z* statistic
5. p value
6. Confidence Interval

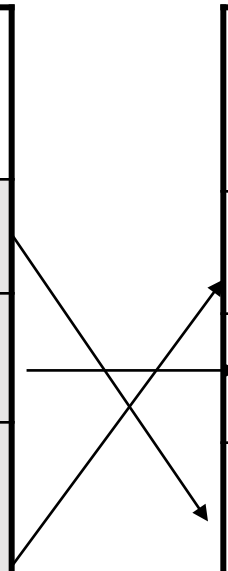
Proportional Odds Model

	Categories			
Cut pt A	0	1	2	3
Cut pt B	0	1	2	3
Cut pt C	0	1	2	3

Proportional Odds Model

Palindromic Invariance

	Categories			
Cut pt A	0	1	2	3
Cut pt B	0	1	2	3
Cut pt C	0	1	2	3



	Categories			
Cut pt A	3	2	1	0
Cut pt B	3	2	1	0
Cut pt C	3	2	1	0

Results reversed (signs) substantive meaning not changed – This can work well with attitude scales!

Continuation Ratio Model

	Categories			
Cut pt A	0	1	2	3
Cut pt B		1	2	3
Cut pt C			2	3

Gayle, V. (1996) 'Modelling Tabular Data with an Ordered Outcome', Sociological Research Online, 1(3). <http://www.socresonline.org.uk/1/3/4.html>

Continuation Ratio Model

	Categories			
Cut pt A	0	1	2	3
Cut pt B		1	2	3
Cut pt C			2	3

	Categories			
Cut pt A	3	2	1	0
Cut pt B		2	1	0
Cut pt C			1	0

Results and substantive meaning are changed

Not Palindromic Invariance

Ordinal (Logit) Models

Ologit model (proportional odds model)

Continuation ratio model

The β that refer to the cut points (or partitions)

in these two ordinal models have slightly different interpretations

Some thoughts on these ordinal models

- Proportional Odds
  - Palindromic invariance (e.g. attitudinal scores)
  - Motivated by an appeal to the existence of an underlying continuous and perhaps unobservable random variable – proportional odds

- Continuation ratio model
  - Natural base line (hierarchy in data)
  - Single direction of movement
  - Categories of Y really are discrete
  - Y categories denotes a shift or change from one state to another not a coarse groupings of some finer scale

Count Data Models

Poisson regression fits models of the number of occurrences (counts) of an event

The Poisson distribution has been applied to diverse events, such as

the number of soldiers kicked to death by horses in the Prussian army (von Bortkiewicz 1898);

the pattern of hits by buzz bombs launched against London during World War II (Clarke 1946);

telephone connections to a wrong number (Thorndike 1926) (Stata Manual)

How many anythings?

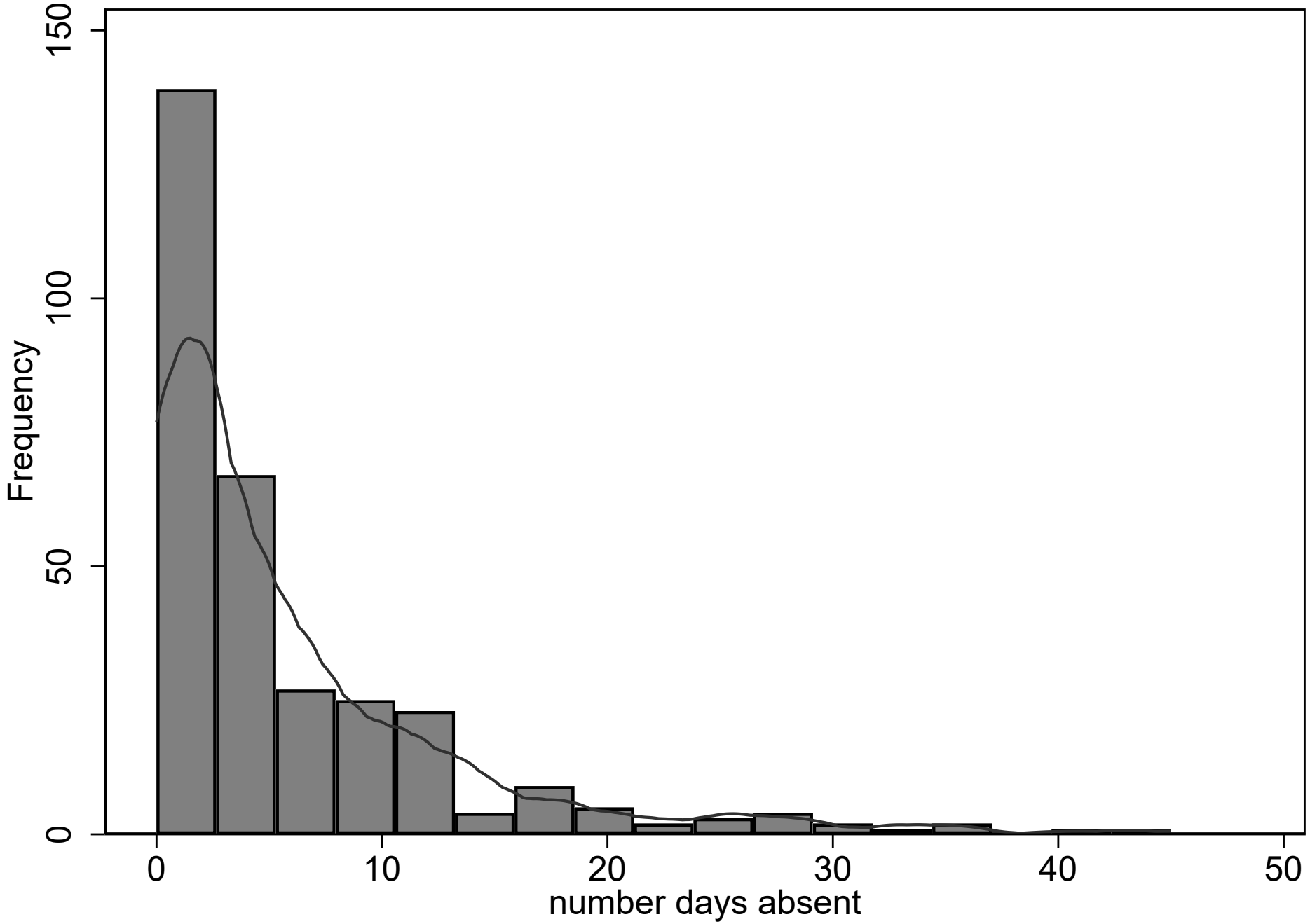
Professor Sir David Spiegelhalter

GP Appointments missed

Beach visits in the last summer

GCSE passes (above grade 4) in school year 11

Number of sexual partners

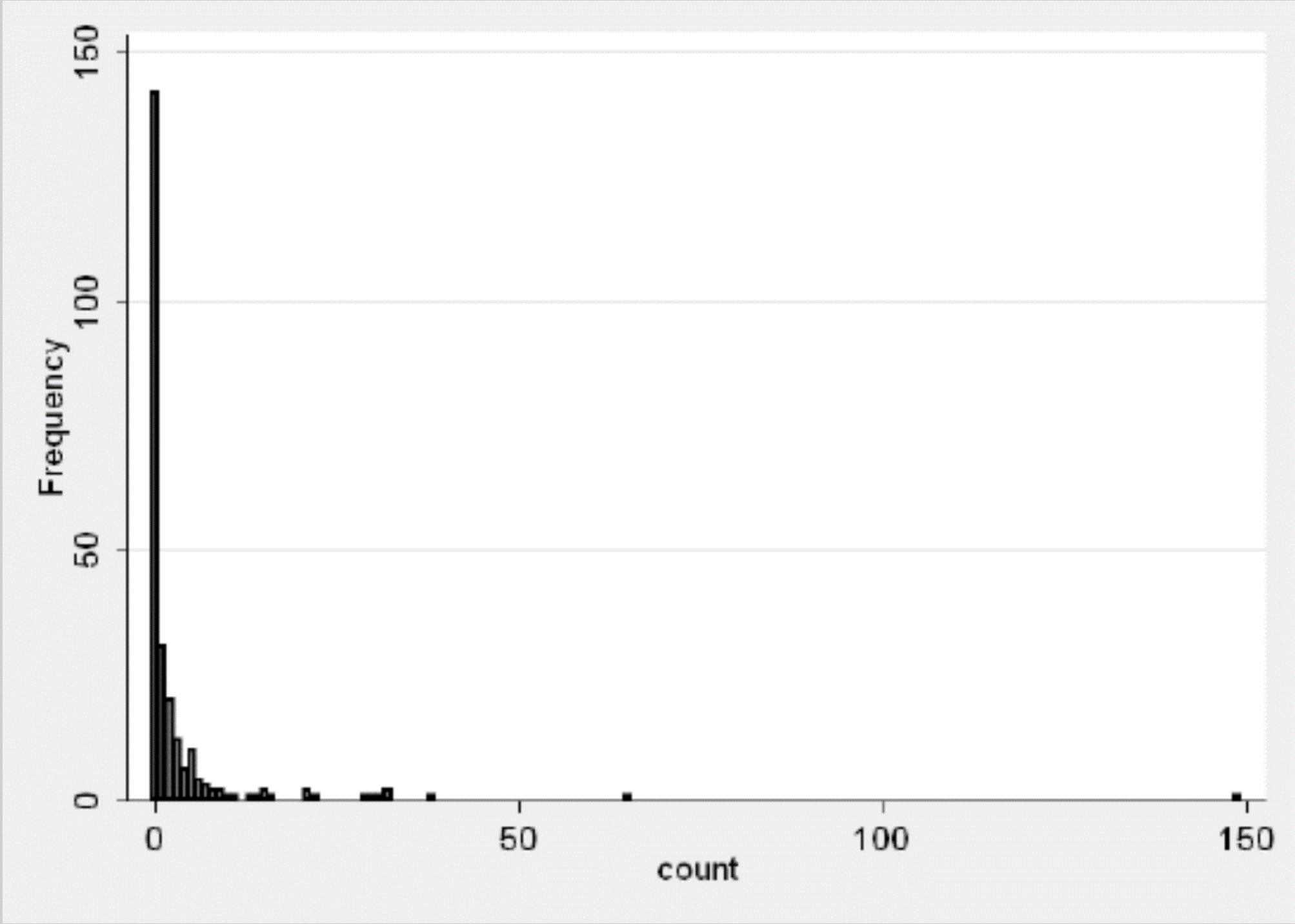

Frequency
0
50
100
150
0
10
20
30
40
50
number days absent

Negative Binomial Model

alpha – The estimate of the dispersion parameter.

If the dispersion parameter, alpha, is significantly greater than zero than the data are over dispersed and are better estimated using a negative binomial

In the poisson model the dispersion parameter is constrained to equals zero


Frequency
0
50
100
150
0
50
100
150
count

Zero Inflated Models

Zero-inflated regression are used to model count data that have an excess of zero counts

In theory excess zeros are generated by a separate process from the count values

The excess zeros can be modelled independently.

The model has two parts, a logit predicating 0 and a count model

Probit

- Binary (0,1)
- Popular in economics
- Largely equivalent to logit model

Y is either 0 or 1

Alternative error structure

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Estimates the effect a one unit change in X_1

$F(\beta_1)$ cumulative distribution function of the standard normal

Probit and Logit

(β_{1logit}) 1.21 / $(\beta_{1probit})$ 0.75 = 1.61

$(\beta_{1probit})$ 0.75 * 1.6 = 1.20

(β_{1logit}) 1.21 / 1.6 = 0.76

Heckman (two-step)

- Probit (0,1) stage 1
- Linear regression (continuous) stage 2

Example:

Does the firm offer overtime?

How many hours of overtime worked

Log-linear Models

- Popular in social stratification research (in sociology)
- Model for multi-way tables (e.g. class/party relationships across elections)
- Tend to model category frequencies (rather then individual outcomes)

The Analytical Problem – Functional Form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

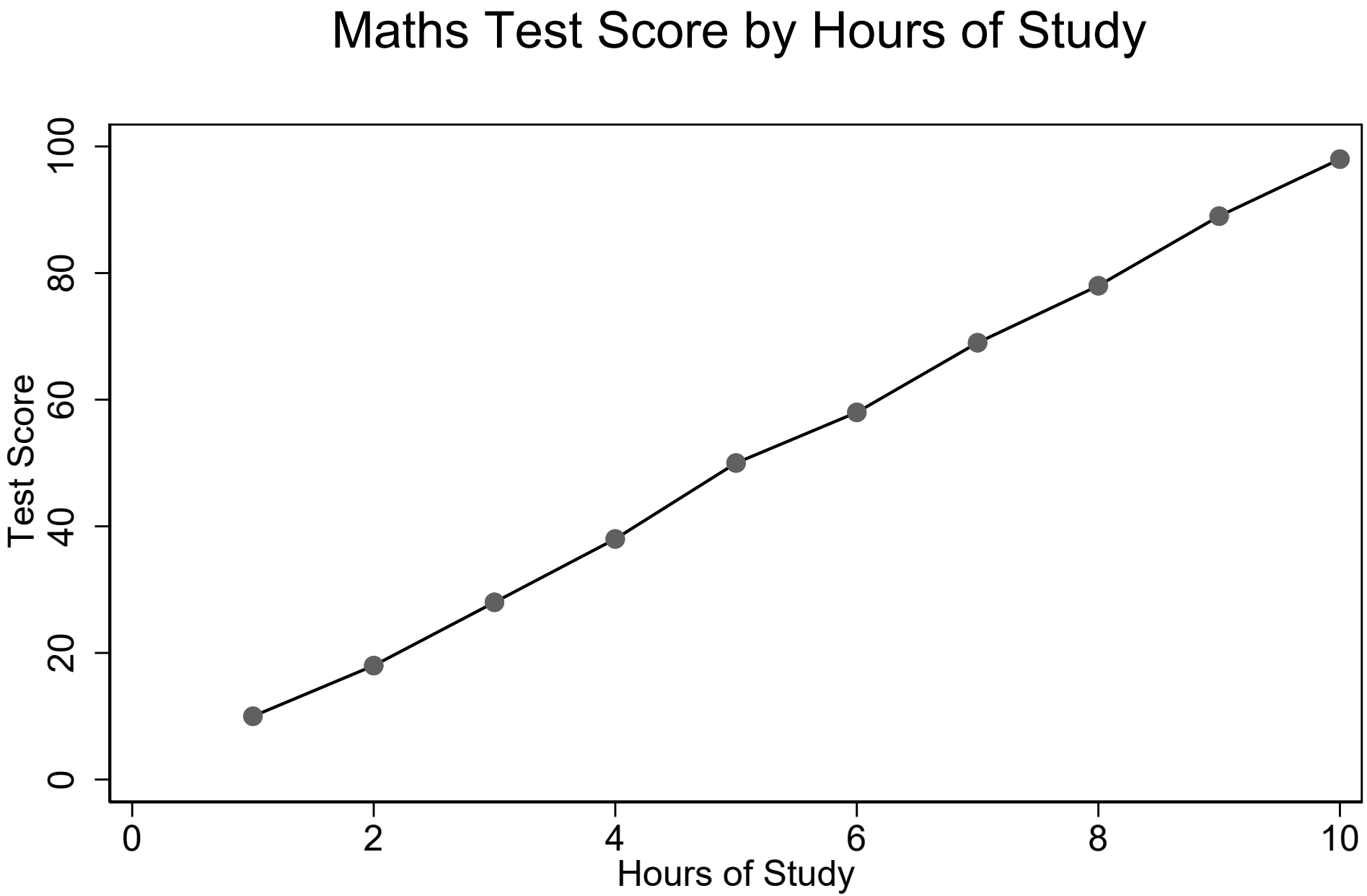
THE UNIVERSITY OF EDINBURGH

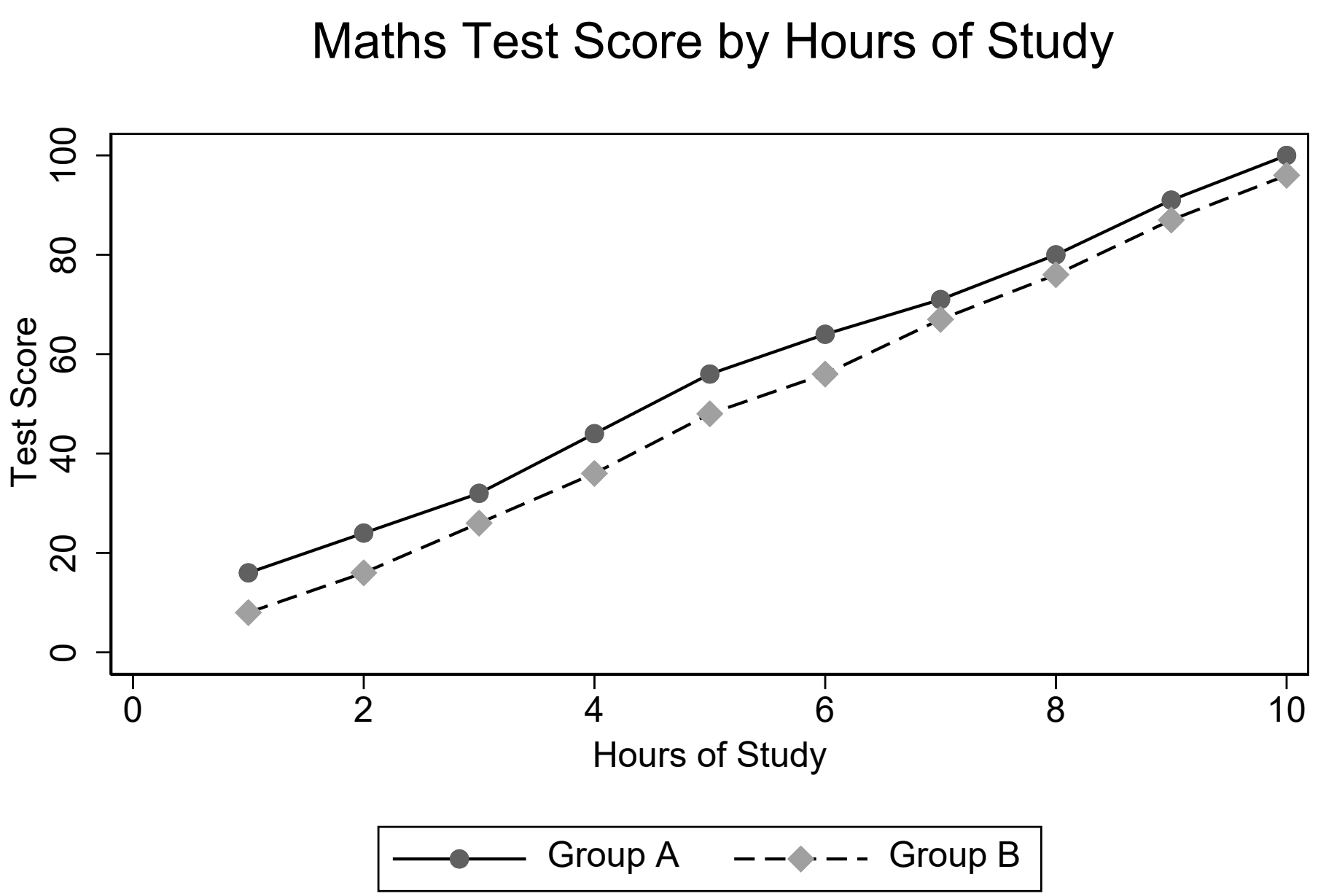
Interaction Effects

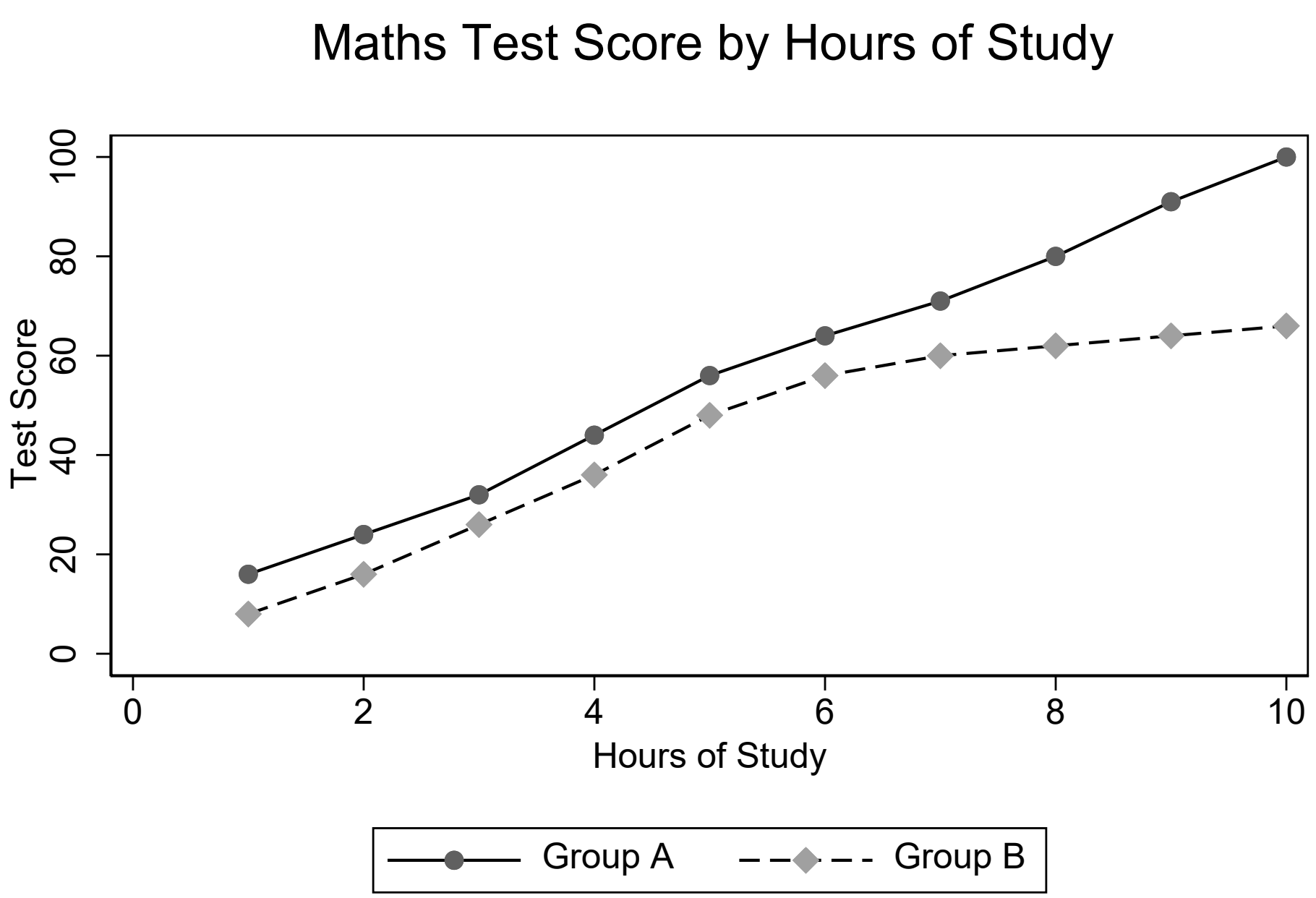
Interaction Effects

‘While sometimes used in the broad sense of effects not operating separately, in statistical discussions it is typically restricted to effects that do not act additively on some response variable’

Oxford Dictionary of Statistical Terms p.203


Maths Test Score by Hours of Study
Test Score
0
20
40
60
80
100
0
2
4
6
8
10
Hours of Study


Maths Test Score by Hours of Study
Test Score
0
20
40
60
80
100
0
2
4
6
8
10
Hours of Study
Group A
Group B


Maths Test Score by Hours of Study
Test Score
0
20
40
60
80
100
0
2
4
6
8
10
Hours of Study
Group A
Group B

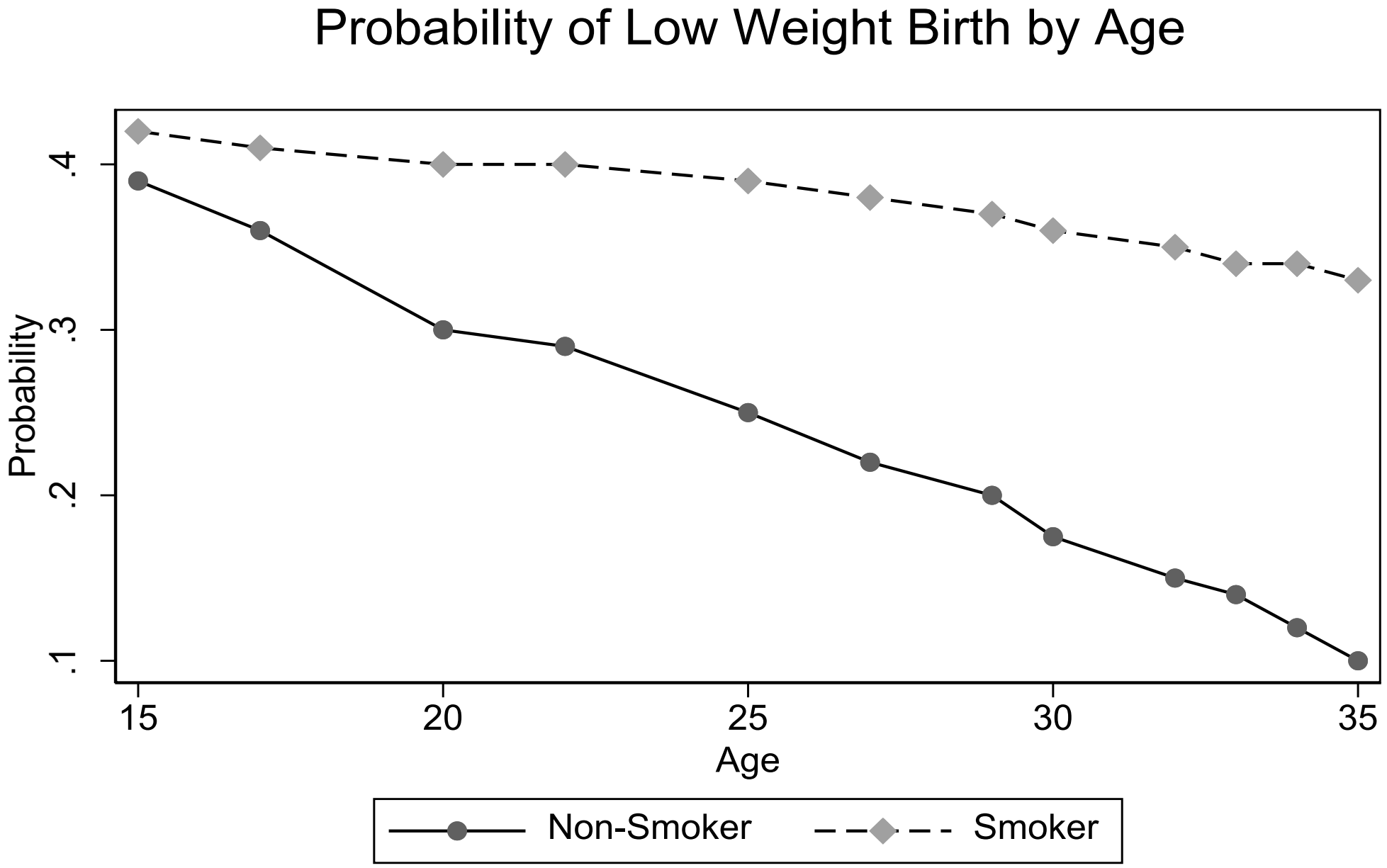
Interaction Effects

An interaction effect – when the effect of

X_2 is also contingent on the value of X_1

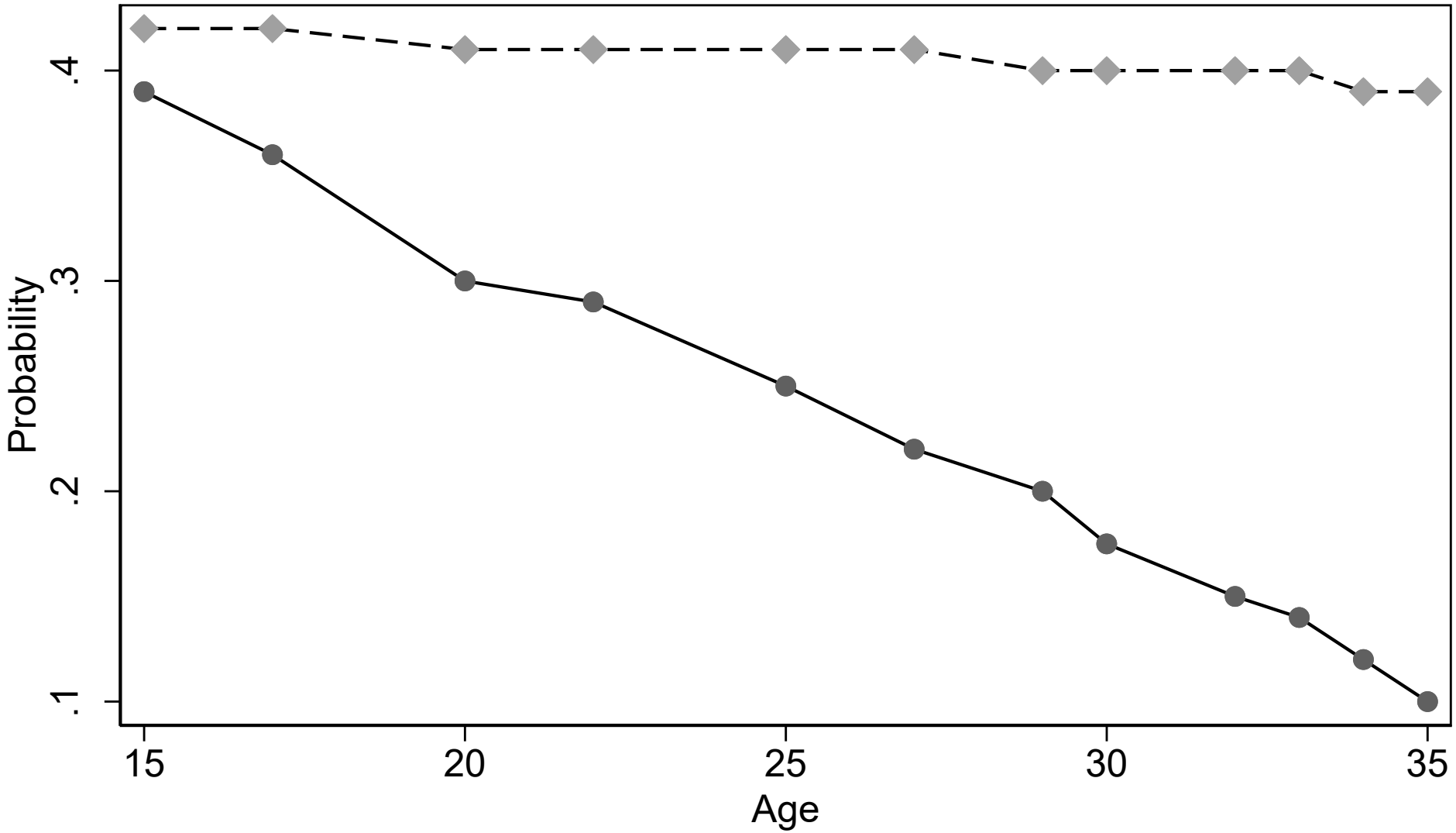
Alternatively….

The effect of X_2 is not uniform across X_1



Example similar to Garrett et al. University of N. Carolina.

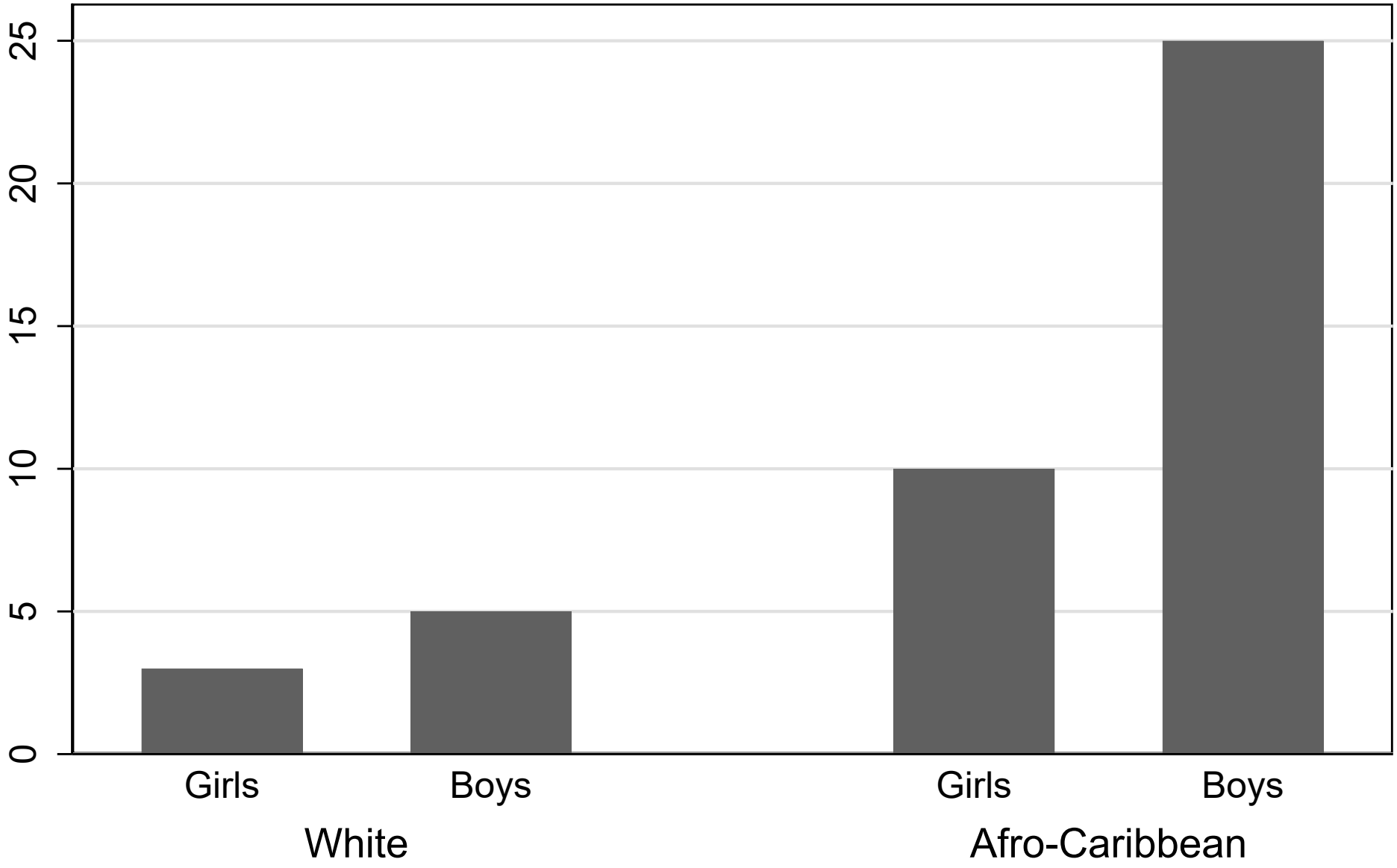
Probability of Low weight Birth by Age



Example similar to Garrett et al. University of N. Carolina.

School Exclusions

Pupils in a London Secondary School



Higher Order Interactions

The effect of X_3 is not uniform across X_2 and X_1

In practice significant interaction effects are less common and main effects will dominate

(despite what awkward audiences at conferences might suggest)

Higher Order Interactions

Historical example
(children respiratory health)

Parent smokes	X_1
Coal fire	X_2
Damp house	X_3

Higher Order Interactions

Two-way Interactions

Parent smokes * Coal fire	$X_1 * X_2$
Parent smokes * Damp house	$X_1 * X_3$
Coal Fire * Damp house	$X_2 * X_3$

Higher Order Interactions

Two-way Interactions

Parent smokes * Coal fire	$X_1 * X_2$
Parent smokes * Damp house	$X_1 * X_3$
Coal Fire * Damp house	$X_2 * X_3$

Parent smokes * Coal fire * Damp house

$X_1 * X_2 * X_3$

Higher Order Interactions

Main effects will tend to be strongest!

In practice it is unusual for higher order interactions to be significant if lower order interactions or main effects are not

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

Statistical Modelling

Software

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

Tools of the Trade

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

Stata https://www.stata.com/

SPSS https://www.ibm.com/uk-en/analytics/spss-statistics-software

SAS https://www.sas.com/

R https://www.r-project.org/

Python https://www.python.org/
(https://www.statsmodels.org/stable/index.html)

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

Considerations

1. Supervisor's expertise
2. Peer group (e.g. other PhD students)
3. Departmental access and support
4. University licenses
5. Data format and meta data (e.g. UK Data Service)
6. Academic subject area
7. Academic job market
8. Non-academic job market

Stata

```
logit admit gre gpa
```

SPSS

```
logistic regression admit with gre gpa.
```

SAS

```
proc logistic data="c:\data\binary" descending;
 class rank / param=ref ;
 model admit = gre gpa;
 run ;
```

R

```
mylogit <- glm(admit ~ gre + gpa, data = mydata, family = "binomial")
```

Python

```
independentVar = [`gre', `gpa',`Int']
logReg = sm.Logit(df[`admit'] , df[independentVar])
answer = logReg.fit( )
```

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

Unified Framework for Modelling

glm

- anova
- Linear regression
- Binary outcomes (logit)
- Categorical models
- Possion regression
- Ordered categorical models

Radical Claim . . .

The glm is simply a special case of the glmm

Generalized Liner Mixed Model glmm

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \upsilon_i + \varepsilon_i$$

LHS = Right Hand Side + Error

The glmm

Unified modelling framework for analysing

1. Hierarchies & Clusters (e.g. multi-levels or geography)
2. Longitudinal data (e.g. panel models)
3. Frailty (e.g. unobservable effect)

Unified Framework for Modelling

glm

- anova
- Linear regression
- Binary outcomes (logit)
- Categorical models
- Possion regression
- Ordered categorical models

glmm

- Clustered data
- Hierarchical data structures
- Repeated contacts (panel)
- Frailty

Unified Framework for Modelling

glm

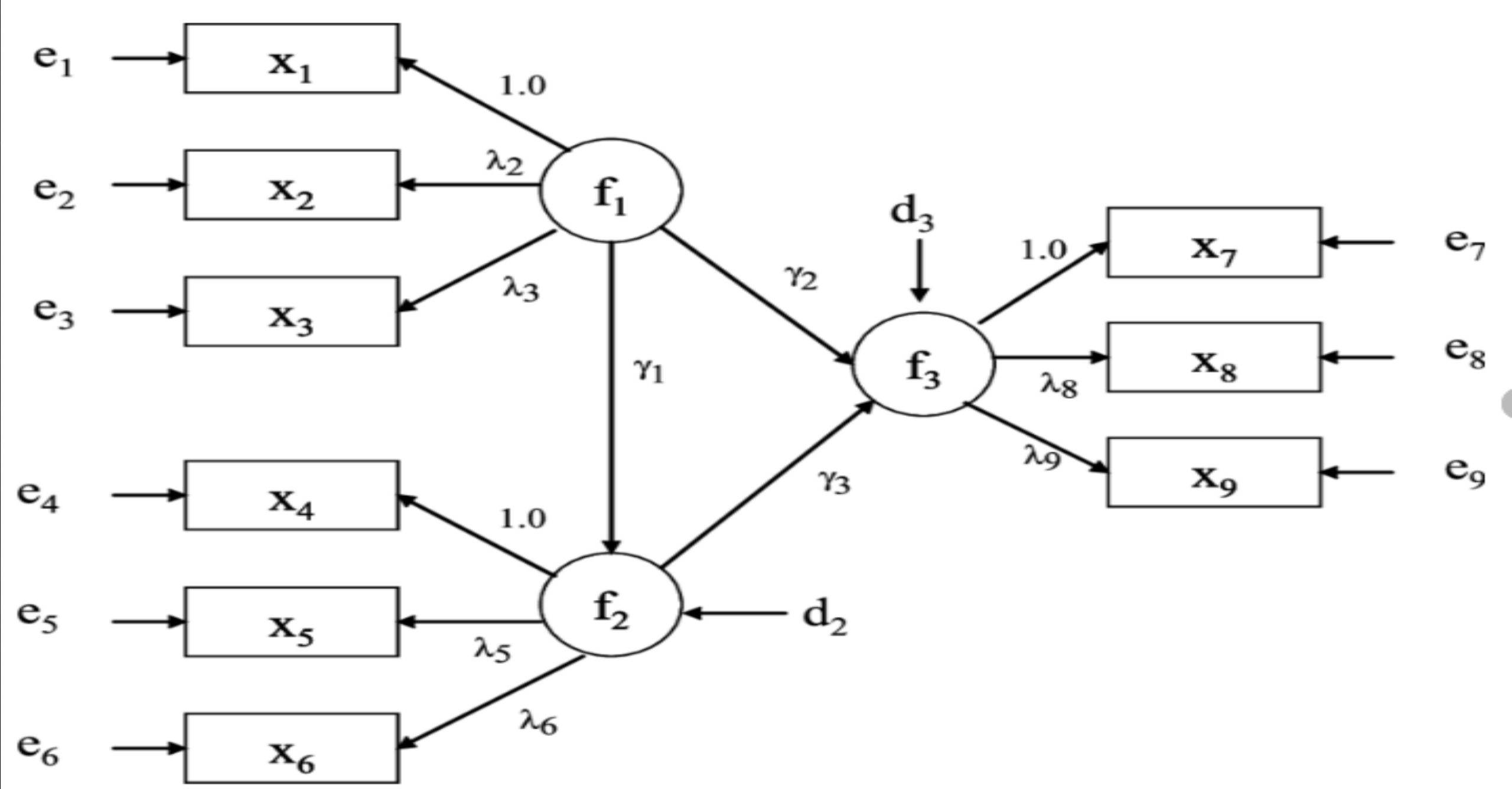
- anova
- Linear regression
- Binary outcomes (logit)
- Categorical models
- Possion regression
- Ordered categorical models

glmm

- Clustered data
- Hierarchical data structures
- Repeated contacts (panel)
- Frailty

gllamm

- Factor analysis
- Latent variables
- Structural Equations



The path diagram for a structural equation model with three factors.

In Your Future Work

- Working with large-scale datasets
- Often nationally representative
- Frequently with complex designs and selection strategies
- Usually testing an idea (detailed empirical work)
- Tends to be a sophisticated description (i.e. a multivariate analysis) rather than an 'causal' analysis

- Data quality
- Appropriately representing survey design and selection
- Undertaking sensitivity analyses
- Thinking about the (potential) effects of missing data
- Maintaining a good workflow (e.g. with electronic research notebooks such as Jupyter)
- Duplicate; Replicate

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

UKRI Economic and Social Research Council

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

NCRM NATIONAL CENTRE FOR RESEARCH METHODS

Many of the ideas and examples presented in this file draw heavily on previous work. We are grateful for the comments and feedback from participants
of earlier workshops.

Professor Vernon Gayle (vernon.gayle@ed.ac.uk)

UKRI Economic and Social Research Council
University of Southampton
Manchester 1824 The University of Manchester
The University of Edinburgh