

## Outline

- Lecture 0: Course Outline
- Lecture 1: Categorical data analysis
- Lecture 2: Ordinal association
- Lecture 3: Multidimensional causality
- Lecture 4: Summary of multiple regression
- Lecture 5: Interaction and Non-linearity
- Lecture 6: Residuals and Influence
- Lecture 7: Logs and log regression
- Lecture 9: Logistic regression
- Lecture 10: Logistic regression continued
- Lecture 11: Multinomial and Ordinal regression

## SO5032 Spring 2023/4 – Module outline

Module Code:	SO5032
Module Title:	Quantitative Research Methods II (MA)
Academic Year:	2023/4
Semester:	Spring
Lecturer(s):	Dr Brendan Halpin
Lecture Locations:	Lec Mon 09-1100 P1006, Lab Weds 12-1400 A0060a
Lecturer(s) Contact Details:	brendan.halpin@ul.ie
Lecturer(s) Office Hours:	Mon 1100-1300

## Short Summary of Module:

Intermediate quantitative research methods for sociology, following on from SO5041.

## Aims and Objectives of Module:

- A continuation of SO5041 – builds on what was learnt there
- A deeper look at methods already covered, especially regression
- Related methods more suited to social science data: methods for categorical and ordinal variables, including logistic regression
- Further use of Stata:
  - Use in a production environment – do-files, logging, reproducibility
  - More complex data handling
  - Further analytic procedures
- Secondary analysis: real research with existing data sets

## Learning Outcomes:

- Deeper understanding of methods for analysis of categorical data
- Understanding of the nature of multivariate causality
- Understanding of the theory and practice of multiple linear regression
- An understanding of some methods for regression with categorical dependent variables
- Deeper understanding of sampling practice and theory
- Practical skills for accessing and analysing large-scale data sets
- An ability to read quantitative social research
- Greater competence in Stata, particularly for handling larger projects

## Course Structure:










One two-hour lecture per week, one two-hour lab per week.

## Detailed outline

- Revisit  $\chi^2$ , look at methods for more complex analysis of categorical (nominal and ordinal) data (chapter 8, Agresti) (1-2 weeks)
- Multivariate causality (chapter 10 from Agresti) (1 week)
- Multiple regression (chapters 11, 14 from Agresti) (3 weeks plus)
- More sampling theory: clusters, strata, weighting (1 week)
- Data sets, data archives and secondary analysis (1 week, ongoing in labs)
- Logistic regression: regression where the dependent variable is binary (or multinomial) rather than continuous (chapter 15 from Agresti) (3 weeks plus)
- Reading statistical research – what gets published and how to read it (1-2 weeks/on-going)

## Lecture topics by week

Week beginning	Topic	Lecture Mon 09-1100	Lab Wed 12-1400
1: Jan 29	Categorical data, association in tables	✓	✓
2: Feb 05	Association in ordinal data	X	✓ (lecture)
3: Feb 12	Understanding multidimensional causality	✓	✓
4: Feb 19	Introducing multiple regression	✓	✓
5: Feb 26	Further multiple regression	✓	✓
6: Mar 04	Multiple regression: residuals & influence	✓	✓
7: Mar 11	Regression with logged dependent variables	✓	✓
8: Mar 18	Introducing logistic regression	X	✓ (lecture)
9: Apr 01	Further logistic regression	X	✓ (lecture)
10: Apr 08	Multinomial regression	✓	✓
11: Apr 15	Multinomial and ordinal regression	✓	✓
12: Apr 22	Ordinal regression continued	✓	✓

<p><b>Texts</b></p> <ul style="list-style-type: none"> <li>• Main text: Agresti, <i>Statistical Methods for the Social Sciences</i> – particularly chapters 8, 10, 11, 14 and 15</li> <li>• Supplementary texts: <ul style="list-style-type: none"> <li>• de Vaus, <i>Surveys in Social Research</i>: good on survey methodology</li> <li>• Agresti, <i>Introduction to Categorical Data Analysis</i></li> <li>• Pevalin and Robson, <i>The Stata Survival Manual</i></li> </ul> </li> </ul> <p>sociology </p>	<p><b>Details of Module Assessment:</b></p> <ul style="list-style-type: none"> <li>• Three assignments, weeks 6, 11 and 15.</li> <li>• The first two assignments are worth 20% each.</li> <li>• The final assignment is a project, worth 60%, and should be worked on throughout the semester (see below).</li> </ul> <p>sociology </p>	<p><b>Details of Annual Repeats:</b></p> <p>A 100% assignment, to be submitted in the examination period.</p> <p>sociology </p>
<p><b>BrightSpace and Other Classroom Technologies:</b></p> <ul style="list-style-type: none"> <li>• The module will use BrightSpace for submission of assignments and for provision of materials.</li> <li>• <a href="http://teaching.sociology.ul.ie/so5032">http://teaching.sociology.ul.ie/so5032</a> will also be used</li> </ul> <p>sociology </p>	<p><b>IN TERM ASSIGNMENT(S):</b></p> <ul style="list-style-type: none"> <li>• Assignment 1: Homework exercises relating to linear regression. <ul style="list-style-type: none"> <li>• Marks: 20%</li> <li>• Deadline: End week 6</li> </ul> </li> <li>• Assignment 2: Homework exercises relating to categorical data analysis. <ul style="list-style-type: none"> <li>• Marks: 20%</li> <li>• Deadline: End week 11</li> </ul> </li> <li>• Assignment 3: A project This will involve the use of large-scale survey data, and require the formulation of a research question, and its addressing using statistical analysis. <ul style="list-style-type: none"> <li>• Marks: 60%</li> <li>• Deadline: End week 15.</li> </ul> </li> </ul> <p>sociology </p>	<p><b>FEEDBACK:</b></p> <p>Detailed feedback on assignments 1 and 2 will be given in weeks 8 and 13, by e-mail and on request face-to-face. Feedback on assignment 3 will be provided on request after the semester.</p> <p>sociology </p>
<p><b>Plagiarism notice</b></p> <p>It hardly needs to be said, but all work must be your own. All material drawn from other sources must be clearly attributed. Passing off others' work as your own is considered academic dishonesty, and can be subject to substantial penalties. Please familiarise yourself with the departmental policy on plagiarism and use the coversheet declaration with all assignments (both available at <a href="http://www.ul.ie/sociology/">http://www.ul.ie/sociology/</a> under Student Resources).</p> <p>sociology </p>	<p><b>Deadline policy</b></p> <p>Please also note the Department's policy on deadlines, also available at <a href="http://www.ul.ie/sociology/">http://www.ul.ie/sociology/</a> under Student Resources.</p> <p>sociology </p>	<p><b>Association between categorical variables</b></p> <ul style="list-style-type: none"> <li>• Association between categorical variables: departure from independence</li> <li>• Visible in patterns of percentages</li> <li>• Three main questions (cf Agresti/Finlay p265) <ul style="list-style-type: none"> <li>• Is there evidence of association?</li> <li>• What is the form of the association?</li> <li>• How strong is the association?</li> </ul> </li> </ul> <p>sociology </p>

## The $\chi^2$ test

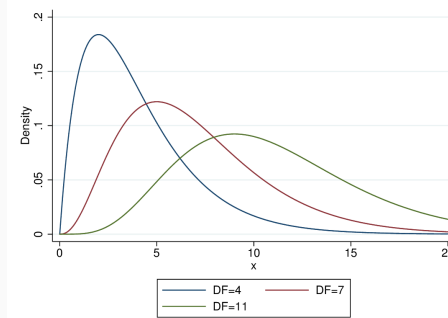
- Compare observed values with expected values under independence:

$$E = \frac{RC}{T}$$

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- For frequency data, and for large samples the  $\chi^2$  statistic has a  $\chi^2$  distribution with  $df = (r - 1)(c - 1)$
- Interpretation: chance of getting a  $\chi^2$  this big or bigger if  $H_0$  (independence) is true in the population

## The $\chi^2$ distribution



## Limitations of $\chi^2$

- Large sample required: most expected counts 5+
- For frequency or count data, not rates or percentages
- Tests for *evidence* of association, not strength (see Agresti/Finlay Table 8.14, p 268)
- Looks for unpatterned association, may miss weak systematic association between ordinal variables

## Pattern of association

- The form association takes is interesting
- We can see it by examining percentages
- Or residuals:  $O - E$
- But residuals depend on sample and expected value size

## Pearson residuals

- "Pearson residuals" are better:

$$\frac{O - E}{\sqrt{E}}$$

- Square and sum these residuals to get the  $\chi^2$  statistic

## Adjusted Residuals

- The sum of squared Pearson residuals has a  $\chi^2$  distribution, but individually they are not normally distributed
- Adjusted residuals scale to have a standard normal distribution if independence holds:

$$AdjRes = \frac{O - E}{\sqrt{E(1 - \pi_r)(1 - \pi_c)}}$$

- Adjusted residuals outside the range -2 to +2 indicate cells with unusual observed values (< 5% chance)
- Adjusted residuals outside the range -3 to +3 indicate cells with very unusual observed values

## Measures of association

- Evidence, pattern, now strength of association
- A number of measures
  - Difference of proportions
  - Odds ratio
  - Risk ratio (ratio of proportions)
- Focus on 2 by 2 pairs, but can be extended to bigger tables

## Difference of proportions

No association

	Favour	Oppose	Total
White	360	240	600
Black	240	160	400
Total	600	400	1000

Maximal association

	Favour	Oppose	Total
White	600	0	600
Black	0	400	400
Total	600	400	1000

## Difference in proportions

- Difference in proportions (i):  $\frac{360}{600} - \frac{240}{400} = 0.6 - 0.6 = 0$
- Difference in proportions (ii):  $\frac{600}{600} - \frac{0}{400} = 1 - 0 = 1$
- Range: -1 through 0 (no association) to +1

Relative risk

- “Relative risk” of ratio or proportions is also popular
- The ratio of two percentages:

$$RR = \frac{n_{11}/n_{1+}}{n_{21}/n_{2+}}$$

where  $n_{1+}$  indicates the row-1 total *etc.*

- Range = 0 through 1 (no association) to  $\infty$

sociology

27

Odds ratios

- Odds differ from proportions/percentages:
  - Percentage:  $\pi_i = \frac{f_i}{\gamma_{total}}$
  - Odds:  $O_i = \frac{f_i}{Total - f_i} = \frac{\pi_i}{1 - \pi_i}$
- Odds ratios are the ratios of two odds:
 
$$OR = \frac{n_{11}/n_{12}}{n_{21}/n_{22}}$$

- Range: 0 though 1 (no association) to  $\infty$

sociology

28

Odds ratios

- Odds ratio (i):  $\frac{\frac{360}{240}}{\frac{240}{160}} = \frac{1.5}{1.5} = 1$
- Odds ratio (ii):  $\frac{\frac{600}{0}}{\frac{0}{200}} = \frac{\infty}{0} = \infty$
- Range: 0 through 1 (no association) to  $+\infty$

sociology

29

Comparing measures

- Difference of proportions is simple and clear
- Ratio of proportions/Relative Risk is also simple
- Odds ratio is less intuitive but turns out to be mathematically more tractable
- DP and RR less consistent across different base levels of “risk”

sociology

30

Ordinal Data

- $\chi^2$  may miss ordinal association
- Symmetric ordinal measures based on concordant and discordant pairs:  $\gamma$  (gamma), Kendall's  $\tau$  (tau).

sociology

31

Lecture 2

Reading (for this and last week):

- Agresti, Chapter 8

sociology

32

Lecture 2

- Expected values, residuals, adjusted residuals in Stata
- Ordinal association
- Association in multi-way tables
- Multivariate causality

sociology

33

Tabular association in Stata

tabchi procedure allows access to

- Percentages
- Expected values
- Residuals
- Adjusted residuals

sociology

34

Ordinal association

- When variables are ordinal, association may be structured
- High values on X are associated with high values on Y, low with low
- Or vice versa for negative association
- Analogous to correlation
- Examine using percentages, adjusted residuals: ordered pattern

sociology

35

Example: row percentages

```

. tab sipram sipram, row

```

Key		Frequency row percentages					
co-habiting is alright	divorce better than unhappy marriage	agree	neither ag	disagree	stronglyd	Total	
strongly agree		2,381	1,228	304	38	19	3,970
		59.87	30.99	7.66	0.96	0.48	100.00
agree		1,462	4,159	667	103	15	6,426
		22.75	64.72	10.40	1.60	0.23	100.00
neither agree, disagree		485	1,803	717	73	13	3,091
		15.69	58.33	23.20	2.36	0.42	100.00
disagree		156	667	302	143	15	1,213
		12.86	53.34	24.77	11.79	1.24	100.00
stronglydisagree		78	143	129	101	50	501
		15.57	28.64	25.76	20.16	9.98	100.00
Total		4,562	7,980	2,089	488	112	15,201
		30.01	52.50	13.74	3.01	0.74	100.00

sociology

36

Example: observed and expected values

```

. tabh1 sipram sipram

```

co-habiting is alright	strongly agree	divorce better than unhappy marriage	agree	neither agree, disagree	disagree	stronglydisagree
strongly agree	2381	1228	304	38	19	
	1191.444	2084.113	545.578	119.614	29.251	
agree	1462	4159	667	103	15	
	1928.519	3373.428	883.094	183.613	47.346	
neither agree, disagree	485	1803	717	73	13	
	927.646	1622.668	424.781	83.131	22.774	
disagree	156	647	252	143	15	
	364.036	636.703	365.697	36.547	8.937	
stronglydisagree	78	143	129	101	50	
	150.386	263.008	61.850	15.096	3.491	

1 cell with expected frequency < 5

Pearson chi(2)(16) = 4.2e+03 Pr = 0.000

Likelihood ratio chi(2)(16) = 3.3e+03 Pr = 0.000

sociology

37

Example: adjusted residuals

```

. tabh1 sipram sipram, adj res

```

co-habiting is alright	strongly agree	divorce better than unhappy marriage	agree	neither agree, disagree	disagree	stronglydisagree
strongly agree	1191.444	2084.113	545.578	119.614	29.251	
	47.925	-31.604	-12.956	-8.616	-2.213	
agree	1928.519	3373.428	883.094	183.613	47.346	
	-16.713	25.829	-9.351	-8.703	-6.210	
neither agree, disagree	927.646	1622.668	424.781	83.131	22.774	
	-19.463	7.277	17.104	-2.373	-2.303	
disagree	364.036	636.703	365.697	36.547	8.937	
	-13.037	0.612	7.418	15.639	2.122	
stronglydisagree	150.386	263.008	61.850	15.096	3.491	
	-7.173	-10.918	7.297	22.491	24.601	

1 cell with expected frequency < 5

Pearson chi(2)(16) = 4.2e+03 Pr = 0.000

Likelihood ratio chi(2)(16) = 3.3e+03 Pr = 0.000

sociology

38

Measures of ordinal association

- Sometimes Pearson's Correlation is used
- Equivalent to scoring the categories linearly and calculating the conventional correlation

```

. corr lopfam1 lopfam1
(observe=15,201)

```

	lopfam1	lopfam1
lopfam1	1.0000	
lopfam1	0.8831	1.0000

sociology

39

Non-linear correlation

- Assumption of equal intervals problematic (but often reasonably OK)
- Spearman's Rank Correlation is a better solution

```

. spearman lopfam1 lopfam1

```

Number of obs = 15201

Spearman's rho = 0.3840

Test of H0: lopfam1 and lopfam1 are independent

Prob > |t| = 0.0000

sociology

40

Truly ordinal measures

- The Gamma statistic ( $\gamma$ ) is truly ordinal
- Counts "concordant" and "discordant" pairs

$$\gamma = \frac{C - D}{C + D}$$

- Range: -1, 0, 1
- Approximately normal for large samples

sociology

41

Gamma in practice

```

. tab lopfam1 lopfam1, gamma

```

co-habiting is alright	strongly agree	divorce better than unhappy marriage	agree	neither ag	disagree	stronglyd	Total
strongly agree	2,381	1,228	304	38	19		3,970
agree	1,462	4,159	667	103	15		6,426
neither agree, disagree	485	1,803	717	73	13		3,091
disagree	156	647	252	143	15		1,213
stronglydisagree	78	143	129	101	50		501
Total	4,562	7,980	2,089	488	112		15,201

gamma = 0.4975 ASR = 0.009

sociology

42

Variants

- Gamma is symmetrical
- Kendall's tau ( $\tau$ ) is also symmetrical, similar logic
- Somer's d also uses  $C + D$  but is asymmetrical: one variable affecting another (takes account of ties)

sociology

43

Multi-way tables

- How do we think in terms of multi-way tables – more than two dimensions?
- Often, in terms of whether the  $A \times B$  relationship is constant across  $C$

sociology

44

### Scouting example

Scout	Delinquent		Total
	Yes	No	
Yes	36	364	400
No	60	340	400
Total	96	704	800

### Scouting example

Low Church Attendance			
Scout	Delinquent		Total
	Yes	No	
Yes	10	40	50
No	40	160	200
Total	50	200	250

Medium Church Attendance			
Scout	Delinquent		Total
	Yes	No	
Yes	18	132	150
No	18	132	150
Total	36	264	300

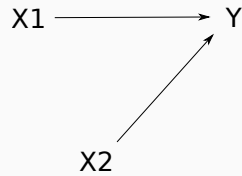
High Church Attendance			
Scout	Delinquent		Total
	Yes	No	
Yes	8	192	200
No	2	48	50
Total	10	240	250

### Multidimensional causality

- Regression analysis never proves causal relationships, but it "thinks" in causal terms
- To use it we need to understand causal relationships: what process generates the data we see, and what can regression tell us about it.
- Start by considering the relationship between variables and patterns of association

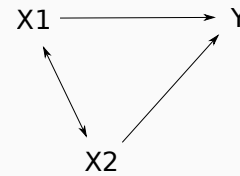
### 3-variable pictures

- Let's consider patterns of causality and association between three variables, X1 and X2, and Y
- If X1 and X2 are not correlated with each other, their separate effects on Y more or less just add up



### Correlated X variables

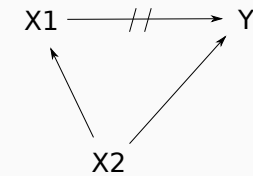
- But if X1 and X2 are correlated, things can get funny:



- In particular, if we measure the effect of one X without taking account of the other we will likely over-estimate it

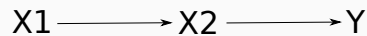
### Spurious association

- X1 may have an association with Y, implying a causal relationship
- But if X2 affects both X1 and Y the relationship between X1 and Y may be **spurious**



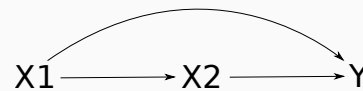
### Indirect effects

- Where there is a time-order (X1 before X2), we may see direct and indirect effects
- X1 may affect X2, which affects Y, but not affect Y directly
- Thus there is association between X1 and Y without a direct causal effect



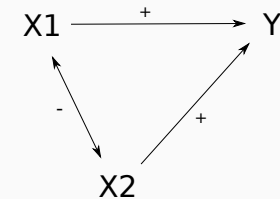
### Direct and indirect effects

- However, it is possible for both direct and indirect effects to be present at the same time



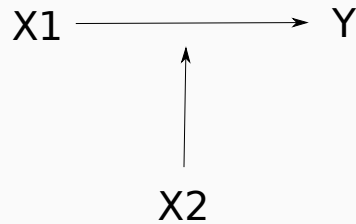
### Suppression

- Where X1 and X2 have positive effects on Y, but a negative correlation, or different effects on Y with a positive correlation, the association between X1 and Y may be **suppressed**
- That is, it may be invisible if we don't take account of X2



## Interactions

- An interaction effect is where the effect of one variable on Y changes depending on the value of another



## Lecture 3: Multidimensional causality

### Multiple regression

## Multiple explanatory variables

- Regression analysis can be extended to the case where there is more than one explanatory variable – multivariate regression
- This allows us to estimate the net simultaneous effect of many variables, and thus to begin to disentangle more complex relationships
- Interpretation is relatively easy: each variable gets its own slope coefficient, standard error and significance
- The slope coefficient is the effect on the dependent variable of a 1 unit change in the explanatory variable, *while taking account of the other variables*

## Example

- Example: income may be affected by gender, and also by paid work time: competing explanations – one or the other, or both could have effects
- We can fit bivariate regressions:

$$\text{Income} = a + b \times \text{PaidWork}$$

or

$$\text{Income} = a + b \times \text{Female}$$

- We can also fit a single multivariate regression

$$\text{Income} = a + b \times \text{PaidWork} + c \times \text{Female}$$

## Dichotomous variables

- We deal with gender in a special way: this is a *binary* or *dichotomous* variable – has two values
- We turn it into a yes/no or 0/1 variable – e.g., female or not
- If we put this in as an explanatory variable a *one-unit change in the explanatory variable* is the difference between being male and female
- Thus the *c* coefficient we get in the  $\text{Income} = a + b \times \text{PaidWork} + c \times \text{Female}$  regression is the net change in predicted income for females, once you take account of paid work time.
- The *b* coefficient is then the net effect of a unit change in paid work time, once you take gender into account.

## Income, hours and gender

```
. corr Income Gender Hours
(obe=506)
```

	Income	Gender	Hours
Income	1.0000		
Gender	-0.3280	1.0000	
Hours	0.3638	-0.4360	1.0000

## Income, hours and gender



## T-test: Income by gender

```
. ttest Income, by(Gender)
Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]
male	437	1618.348	59.11677	1235.809	1502.159 1734.537
female	531	992.1805	40.82127	940.6625	911.9692 1072.372
combined	968	1274.861	36.23219	1127.261	1203.759 1345.964
diff		626.1674	70.00484		488.7883 763.5465

diff = mean(male) - mean(female)      t = 8.9446  
H0: diff = 0      degrees of freedom = 966  
Ha: diff < 0      Ha: diff != 0      Ha: diff > 0  
Pr(T < t) = 1.0000      Pr(|T| > |t|) = 0.0000      Pr(T > t) = 0.0000

## Regression: Just hours

```
. reg Income Hours
```

Source	SS	df	MS	Number of obs	=	506
Model	86947928.8	1	86947928.8	F(1, 504)	=	76.88
Residual	570128215	504	1131206.78	Prob > F	=	0.0000
				R-squared	=	0.1323
				Adj R-squared	=	0.1306
				Root MSE	=	1063.6
Total	657076144	505	1301140.88			

Income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Hours	37.82204	4.314061	8.77	0.000	29.34628 46.2978
_cons	449.7435	150.1722	2.99	0.003	154.703 744.7841

## Regression: Hours and binary gender

```
. reg Income Hours i.Gender
```

Source	SS	df	MS	Number of obs	=	506
Model	110256231	2	55118115.6	F(2, 503)	=	50.70
Residual	546839912	503	1087166.88	Prob > F	=	0.0000
				R-squared	=	0.1678
				Adj R-squared	=	0.1645
Total	657076144	505	1301140.88	Root MSE	=	1042.7

	Income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Hours		28.33857	4.699451	6.03	0.000	19.1056 37.57155
Gender						
female		-478.4214	103.3684	-4.63	0.000	-681.5084 -275.3344
_cons		1022.139	192.2717	5.32	0.000	644.3844 1399.893

## Regression: for men only

```
. reg Income Hours if Gender==1
```

Source	SS	df	MS	Number of obs	=	232
Model	8009519.02	1	8009519.02	F(1, 230)	=	5.36
Residual	343845612	230	1494980.92	Prob > F	=	0.0215
				R-squared	=	0.0228
				Adj R-squared	=	0.0185
Total	351855131	231	1523182.38	Root MSE	=	1222.7

	Income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Hours		24.61855	10.63597	2.31	0.022	3.662162 45.57495
_cons		1164.366	414.4901	2.81	0.005	347.6826 1981.049

## Regression: for women only

```
. reg Income Hours if Gender==2
```

Source	SS	df	MS	Number of obs	=	274
Model	31772944.2	1	31772944.2	F(1, 272)	=	42.63
Residual	202744304	272	745383.469	Prob > F	=	0.0000
				R-squared	=	0.1355
				Adj R-squared	=	0.1323
Total	234517248	273	859037.537	Root MSE	=	863.36

	Income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Hours		29.70376	4.549594	6.53	0.000	20.74687 38.66065
_cons		504.6153	140.3614	3.60	0.000	228.2824 780.9482

## Regression: interaction

```
. reg Income c.Hours#i.Gender
```

Source	SS	df	MS	Number of obs	=	506
Model	110486228	3	36828742.8	F(3, 502)	=	33.82
Residual	546599915	502	1088824.53	Prob > F	=	0.0000
				R-squared	=	0.1681
				Adj R-squared	=	0.1632
Total	657076144	505	1301140.88	Root MSE	=	1043.5

	Income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Hours		24.61855	9.076915	2.71	0.007	6.785132 42.45198
Gender						
female		-659.7502	392.3082	-1.68	0.093	-1430.518 111.0181
Gender#c.Hours						
female		5.085207	10.61255	0.48	0.632	-15.76529 25.9357
_cons		1164.366	353.7327	3.29	0.001	469.3865 1859.345

## Regression: Direct and indirect 1

```
. reg ownscore fatherscore
```

Source	SS	df	MS	Number of obs	=	1,000
Model	13269.3853	1	13269.3853	F(1, 998)	=	53.50
Residual	247825.861	998	248.021905	Prob > F	=	0.0000
				R-squared	=	0.0509
				Adj R-squared	=	0.0499
Total	260795.247	999	261.056303	Root MSE	=	15.749

	ownscore	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
fatherscore		.2370829	.032413	7.31	0.000	.1734773 .3006884
_cons		37.90861	1.672157	22.67	0.000	34.62726 41.18996

## Regression: Direct and indirect 2

```
. reg education fatherscore
```

Source	SS	df	MS	Number of obs	=	1,000
Model	311.104929	1	311.104929	F(1, 998)	=	111.01
Residual	2797.00607	998	2.80261129	Prob > F	=	0.0000
				R-squared	=	0.1001
				Adj R-squared	=	0.0992
Total	3108.111	999	3.11122222	Root MSE	=	1.6741

	education	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
fatherscore		.0363018	.0034455	10.54	0.000	.0295405 .0430631
_cons		1.295213	.1777516	7.29	0.000	.9464035 1.644023

## Regression: Direct and indirect 3

```
. reg ownscore education
```

Source	SS	df	MS	Number of obs	=	1,000
Model	80742.8091	1	80742.8091	F(1, 998)	=	447.54
Residual	180052.437	998	180.413264	Prob > F	=	0.0000
				R-squared	=	0.3096
				Adj R-squared	=	0.3089
Total	260795.247	999	261.056303	Root MSE	=	13.432

	ownscore	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
education		5.096871	.2409273	21.16	0.000	4.624089 5.569653
_cons		33.87079	.8556481	39.68	0.000	32.19171 35.54986

## Regression: Direct and indirect 4

```
. reg ownscore education fatherscore
```

Source	SS	df	MS	Number of obs	=	1,000
Model	81453.7212	2	40726.8606	F(2, 997)	=	226.41
Residual	179341.525	997	179.881169	Prob > F	=	0.0000
				R-squared	=	0.3123
				Adj R-squared	=	0.3109
Total	260795.247	999	261.056303	Root MSE	=	13.412

	ownscore	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
education		4.937369	.2635982	19.47	0.000	4.439722 5.435017
fatherscore		.0578475	.0290984	1.99	0.047	.0007463 .1149486
_cons		31.51367	1.461439	21.56	0.000	28.64592 34.38152

## Formula for multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_k X_k + e$$

$$e \sim N(0, \sigma)$$

- Interpretation of  $\beta_j$ 
  - How much  $\hat{Y}$  changes for a 1-unit in  $X_j$  holding all other values constant
  - The estimated effect on  $Y$  of a 1-unit change in  $X_j$ , "controlling for" or "taking account" of all the other  $X$ s



## Predictions

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_k X_k$$

- Enter values for all X variables to get a prediction for those values
- If we increase  $X_i$  by 1, holding all others the same,  $\hat{Y}$  changes by  $\beta_i$

## Simplest example

- Simplest multiple regression model adds a binary variable to a model with a continuous X

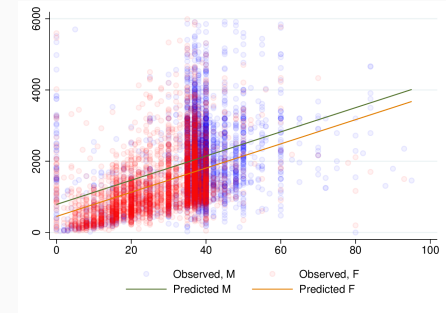
```
. reg income hours i.sex
```

Source	SS	df	MS	Number of obs	=	7,945
Model	1.8935e+09	2	946761687	F(2, 7942)	=	794.96
Residual	9.4586e+09	7,942	1190962.07	Prob > F	=	0.0000
				R-squared	=	0.1668
				Adj R-squared	=	0.1666
				Root MSE	=	1091.3
Total	1.1352e+10	7,944	1429021.17			

	income	Coefficient	Std. err.	t	P> t	[95% conf. interval]
hours		33.96065	1.123629	30.22	0.000	31.75804 36.16326
sex						
female		-337.0889	26.44232	-12.75	0.000	-388.9228 -285.255
_cons		787.1759	45.73595	17.21	0.000	697.5214 876.8304

## Predicted lines: one for each value of sex



## More general 2 X-variable example

```
. reg wage ttl_exp grade
```

Source	SS	df	MS	Number of obs	=	2,244
Model	11010.6	2	5505.3	F(2, 2241)	=	194.77
Residual	63343.7305	2,241	28.2658325	Prob > F	=	0.0000
				R-squared	=	0.1481
				Adj R-squared	=	0.1473
				Root MSE	=	5.3166
Total	74354.3305	2,243	33.1495009			

	wage	Coefficient	Std. err.	t	P> t	[95% conf. interval]
ttl_exp		.2616056	.0248373	10.53	0.000	.2128992 .310312
grade		.6483343	.045426	14.27	0.000	.5592528 .7374158
_cons		-4.002059	.6245962	-6.41	0.000	-5.226906 -2.777211

## Effect of experience on wage, controlling for grade



## Effect of grade on wage, controlling for experience



See <https://teaching.sociology.ul.ie/so5032/ttlgrade.html>

## Residuals

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_k X_k$$

$$Y = \hat{Y} + e$$

$$e \sim N(0, \sigma)$$

- Mean of zero
- Standard deviation of  $\sigma$  (RMSE)
- Normally distributed
- Should have no structured relationship to X variables

## Lecture 4: Summary of multiple regression

### $R^2$

## $R^2$

- $R^2$ : coefficient of multiple determination
- TSS = sum of squared deviation from the mean =  $\sum (Y_i - \bar{Y})^2$
- RSS = sum of squared deviation from the regression prediction =  $\sum (Y_i - \hat{Y}_i)^2$
- $R^2 = \frac{TSS - RSS}{TSS}$
- Range: 0 (no relationship) to 1 (perfect linear relationship)
- PRE: Proportional Reduction in Error

## R<sup>2</sup> and correlation

- In bivariate regression,  $R^2$  is the square of the correlation coefficient between Y and X
- In multiple regression, it is the square of the correlation between Y and  $\hat{Y}$
- (In bivariate regression the correlation between X and  $\hat{Y}$  is 1)

## Lecture 4: Summary of multiple regression

### Hypothesis testing

## Hypothesis testing: one parameter at a time

- t-test:  $abs(\hat{\beta}_j / se_j) > t$
- Interpretation:
  - Null: population value of  $\beta$  is 0; this variable has no influence once the other variables are taken account of

## Example

```
. reg income age i.sex
```

Source	SS	df	MS	Number of obs	=	959
Model	33922983.9	2	16961492	F(2, 956)	=	45.72
Residual	354670636	956	370994.389	Prob > F	=	0.0000
				R-squared	=	0.0873
				Adj R-squared	=	0.0854
				Root MSE	=	609.09
Total	388593620	958	405630.083			

	income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age		-3.144945	1.083398	-2.90	0.004	-5.271057 -1.018833
sex						
female		-352.678	39.51326	-8.93	0.000	-430.2208 -275.1353
_cons		1035.878	54.58935	18.98	0.000	928.7494 1143.007

## Hypothesis testing: all parameters together

- F-test:
  - $\beta_1 = \beta_2 \dots = \beta_k = 0$
- Null hypothesis: no X variable has an effect once the others are taken care of.
- A "global" test: the null is that there is no relevant variable in the model
- Calculation based on TSS and RSS, but also number of cases and number of parameters estimated
- Uses F distribution (two df parameters: k and n-k-1, k is number of parameters, n the number of cases)

## Hypothesis testing: additional parameters

- Delta F-test compares "nested" models
  - Model 1:  $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_g X_g$
  - Model 1:  $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_g X_g + \beta_h X_h \dots + \beta_k X_k$
- Null hypothesis:  $\beta_h = \dots = \beta_k = 0$
- That is, given the variables already in the model, the additional variables contribute no explanatory power.
- Useful when adding multi-category variables, or related groups of variables

## Dummy variables

In regression models we often use "indicator coding" or "dummy coding"

With a two-category variable, we set one category to 0 and the other to 1 and interpret it as the effect of being in the second category (e.g., female) compared with the first.

```
. reg income age i.sex
```

Source	SS	df	MS	Number of obs	=	959
Model	33922983.9	2	16961492	F(2, 956)	=	45.72
Residual	354670636	956	370994.389	Prob > F	=	0.0000
				R-squared	=	0.0873
				Adj R-squared	=	0.0854
				Root MSE	=	609.09
Total	388593620	958	405630.083			

	income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age		-3.144945	1.083398	-2.90	0.004	-5.271057 -1.018833
sex						
female		-352.678	39.51326	-8.93	0.000	-430.2208 -275.1353
_cons		1035.878	54.58935	18.98	0.000	928.7494 1143.007

## More than two categories

With more than two categories we create a set of binary variables, "indicator variables" or "dummy variables":

	d1	d2	d3	d4
a	1	0	0	0
b	0	1	0	0
c	0	0	1	0
d	0	0	0	1

For m categories, m-1 dummy variables are sufficient.

We interpret the parameter as the estimated effect of being in that category relative to the omitted or "reference" category.

Stata handles this automatically with the `i.` prefix.

## Example

```
. reg income age i.sex i.qual
```

Source	SS	df	MS	Number of obs	=	959
Model	85960604.5	5	17192120.9	F(5, 953)	=	54.14
Residual	302633015	953	317568.253	Prob > F	=	0.0000
				R-squared	=	0.2212
				Adj R-squared	=	0.2171
				Root MSE	=	563.52
Total	388593620	958	405630.083			

	income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
age		-3.3897295	1.04777	-0.37	0.710	-2.445933 1.666474
sex						
female		-336.9623	36.75947	-9.17	0.000	-409.1011 -264.8234
qual						
A-level, other sub-d...		-459.9208	78.54165	-5.86	0.000	-614.0554 -305.7862
D-level, commercial...		-701.695	77.16016	-9.09	0.000	-853.1185 -550.2716
Sub-D-level, no qual		-864.9695	76.41768	-11.32	0.000	-1014.936 -715.0032
_cons		1563.508	61.83797	19.10	0.000	1402.904 1724.111

Interactions

- An interaction effect is where the effect of one variable on Y changes depending on the value of another

sociology

87

Income, hours and gender

```
. reg income hours i.sex
```

Source	SS	df	MS	Number of obs	=	7,945
Model	1.8935e+09	2	946761687	F(2, 7942)	=	794.96
Residual	9.4586e+09	7,942	1190962.07	Prob > F	=	0.0000
				R-squared	=	0.1668
				Adj R-squared	=	0.1666
Total	1.1352e+10	7,944	1429021.17	Root MSE	=	1091.3

income	Coefficient	Std. err.	t	P> t	[95% conf. interval]
hours	33.96065	1.123629	30.22	0.000	31.75804 36.16326
sex					
female	-337.0889	26.44232	-12.75	0.000	-388.9228 -285.255
_cons	787.1759	45.73595	17.21	0.000	697.5214 876.8304

sociology

88

For men

```
. reg income hours if sex==1
```

Source	SS	df	MS	Number of obs	=	3,818
Model	344180174	1	344180174	F(1, 3816)	=	204.70
Residual	6.4162e+09	3,816	1681398.47	Prob > F	=	0.0000
				R-squared	=	0.0509
				Adj R-squared	=	0.0507
Total	6.7604e+09	3,817	1771128.3	Root MSE	=	1296.7

income	Coefficient	Std. err.	t	P> t	[95% conf. interval]
hours	28.71923	2.007313	14.31	0.000	24.78372 32.65474
_cons	983.9722	78.23438	12.58	0.000	830.587 1187.357

sociology

89

For women

```
. reg income hours if sex==2
```

Source	SS	df	MS	Number of obs	=	4,127
Model	764315243	1	764315243	F(1, 4125)	=	1043.34
Residual	3.0218e+09	4,125	732568.614	Prob > F	=	0.0000
				R-squared	=	0.2019
				Adj R-squared	=	0.2017
Total	3.7862e+09	4,126	917634.7	Root MSE	=	855.9

income	Coefficient	Std. err.	t	P> t	[95% conf. interval]
hours	38.11874	1.180121	32.30	0.000	35.80507 40.43241
_cons	330.7275	36.40158	9.09	0.000	259.3607 402.0942

sociology

90

Different effects

sociology

91

Interaction in regression

- We can capture interaction effects with a regression model of this form:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

- That is, a 1-unit increase in  $X_1$  leads to a  $\beta_1 + \beta_3 X_2$  increase in  $\hat{Y}$
- Equivalently, a 1-unit increase in  $X_2$  leads to a  $\beta_2 + \beta_3 X_1$  increase in  $\hat{Y}$

sociology

92

Interaction between hours and sex

- Simplest example: one variable is binary

$$\hat{Y}_m = \beta_0 + \beta_1 X_1 + \beta_2 \times 0 + \beta_3 X_1 \times 0$$

$$\hat{Y}_f = \beta_0 + \beta_1 X_1 + \beta_2 \times 1 + \beta_3 X_1 \times 1$$

sociology

93

One-unit increase

If  $X_1$  increases by 1 unit,  $\hat{Y}$  changes:

$$\Delta \hat{Y}_m = \beta_1$$

$$\Delta \hat{Y}_f = \beta_1 + \beta_3$$

sociology

94

Stata: by hand

- First create an interaction variable:

```
gen female = sex == 2
gen intvar = hours*female
```
- Then fit the regression:

```
reg income hours female intvar
```

sociology

95

Results

```

.gen female = sex==2
.gen intvar = female*hours
.reg income hours female intvar

```

Source	SS	df	MS	Number of obs	=	7,945
Model	1.9141e+09	3	638027348	P(3, 7941)	=	536.82
Residual	9.4381e+09	7,941	1188523.12	Prob > F	=	0.0000
				R-squared	=	0.1686
				Adj R-squared	=	0.1683
Total	1.1352e+10	7,944	1429021.17	Root MSE	=	1090.2

	income	Coefficient	Std. err.	t	P> t	[95% conf. interval]
hours		28.71923	1.687655	17.02	0.000	25.41098 32.02747
female		-653.2448	80.47524	-8.12	0.000	-810.9974 -495.4921
intvar		9.399515	2.260017	4.16	0.000	4.969287 13.82974
_cons		983.9722	65.7758	14.96	0.000	855.0344 1112.91

sociology
96

Stata's formula syntax

- But more convenient to use Stata's formula syntax

```
reg income c.hours##i.sex
```

- i.sex means treat sex as categorical
- c.hours#i.sex creates the interaction between hours (continuous, c.) and sex
- c.hours##i.sex puts both the interaction and the first order terms in the model

sociology
97

Same results using Stata's formula syntax

```

. reg income c.hours##i.sex

```

Source	SS	df	MS	Number of obs	=	7,945
Model	1.9141e+09	3	638027348	F(3, 7941)	=	536.82
Residual	9.4381e+09	7,941	1188523.12	Prob > F	=	0.0000
				R-squared	=	0.1686
				Adj R-squared	=	0.1683
Total	1.1352e+10	7,944	1429021.17	Root MSE	=	1090.2

	income	Coefficient	Std. err.	t	P> t	[95% conf. interval]
hours		28.71923	1.687655	17.02	0.000	25.41098 32.02747
sex						
female		-653.2448	80.47524	-8.12	0.000	-810.9974 -495.4921
sex#c.hours						
female		9.399515	2.260017	4.16	0.000	4.969287 13.82974
_cons		983.9722	65.7758	14.96	0.000	855.0344 1112.91

sociology
98

Predictions

Sex	Hrs	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\hat{y}$
M	0	983.9722	+ 0*28.71923	+ 0*-653.2448	+ 0*0*9.399515	= 983.9722
M	80	983.9722	+ 80*28.71923	+ 0*-653.2448	+ 80*0*9.399515	= 3281.5106
F	0	983.9722	+ 0*28.71923	+ 1*-653.2448	+ 0*1*9.399515	= 330.7274
F	80	983.9722	+ 80*28.71923	+ 1*-653.2448	+ 80*1*9.399515	= 3380.227

sociology
99

Interactions between two continuous variable

```

. reg wage c.ttl_exp#c.grade

```

Source	SS	df	MS	Number of obs	=	2,244
Model	11301.2662	3	3767.08872	F(3, 2240)	=	133.83
Residual	63053.0643	2,240	28.1486894	Prob > F	=	0.0000
				R-squared	=	0.1520
				Adj R-squared	=	0.1509
Total	74354.3305	2,243	33.1495009	Root MSE	=	5.3055

	wage	Coefficient	Std. err.	t	P> t	[95% conf. interval]
ttl_exp		-.143543	.1284932	-1.12	0.264	-.3955211 .1084352
grade		.2515455	.1315367	1.91	0.056	-.0064011 .5094921
c.ttl_exp#c.grade		.032074	.0099813	3.21	0.001	.0125005 .0510475
_cons		.933757	1.657647	0.56	0.573	-2.316929 4.184443

sociology
100

Without interaction, predictions for different levels of grade

Wage predicted by work experience and tenure, no interaction

sociology
101

With interaction

Interaction between work experience and tenure

sociology
102

Lecture 5: Interaction and Non-linearity

Non-linear linear regression

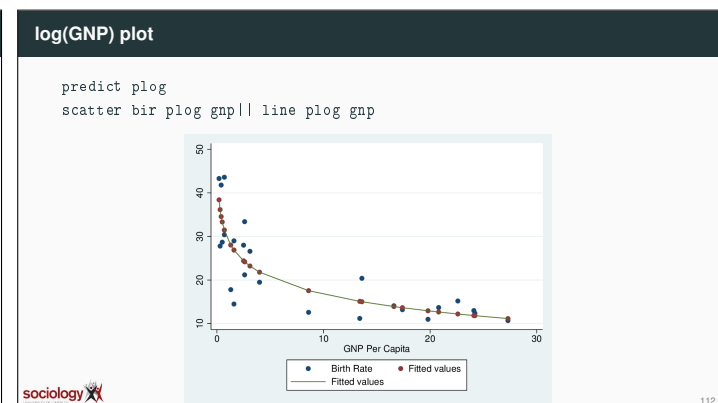
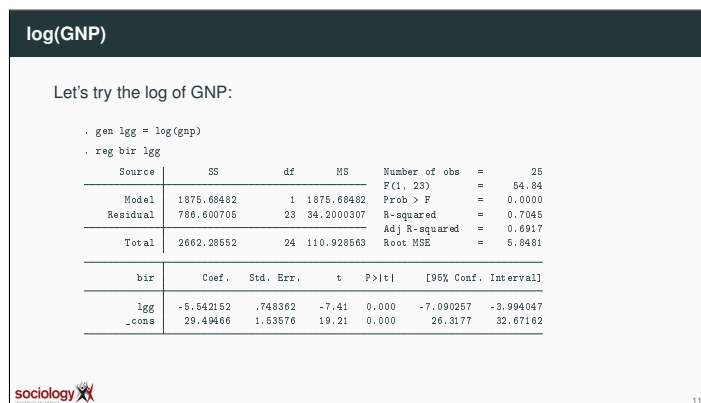
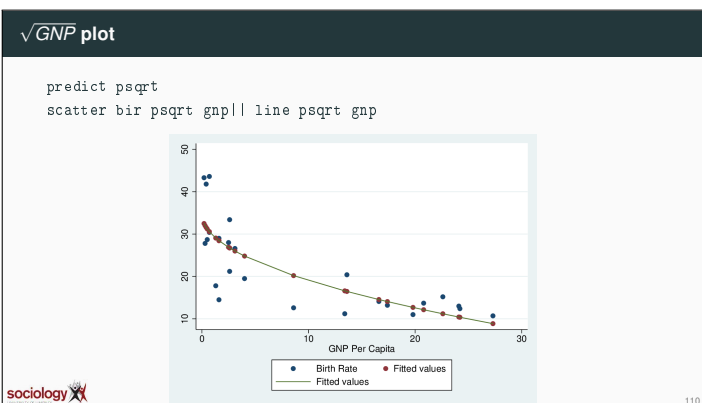
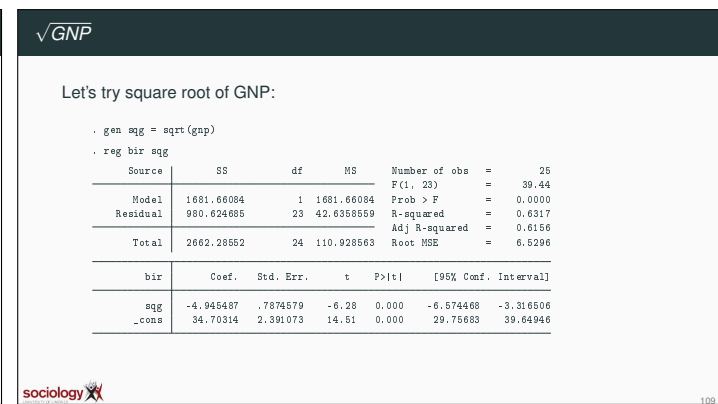
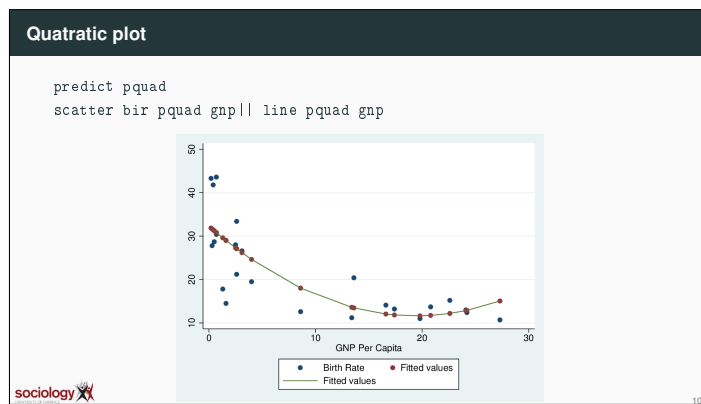
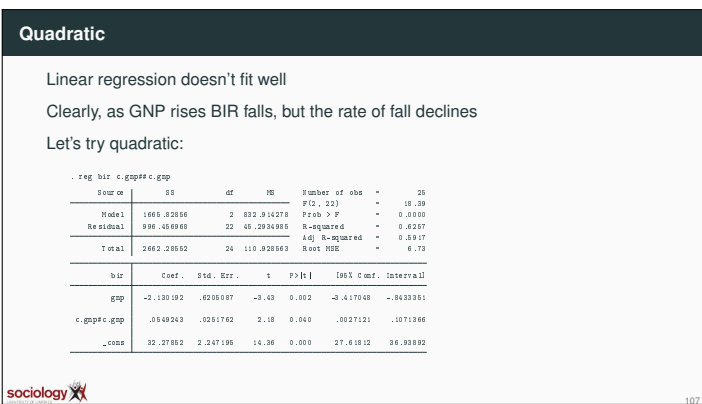
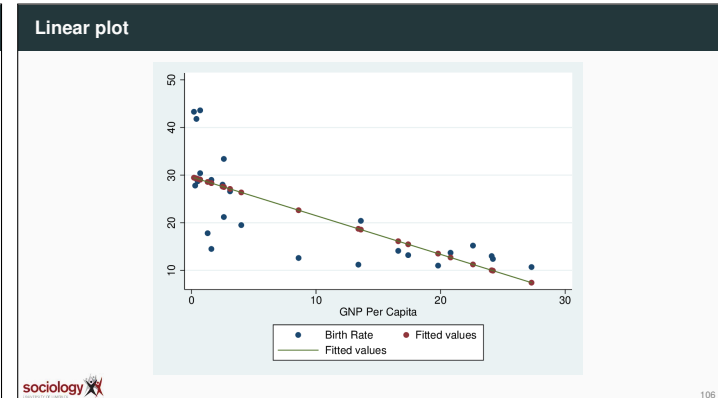
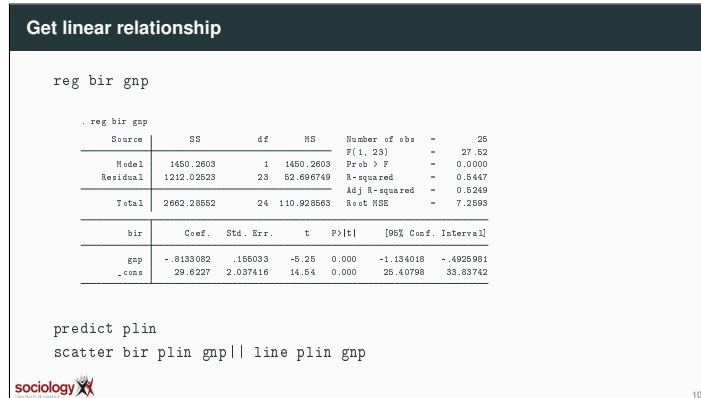
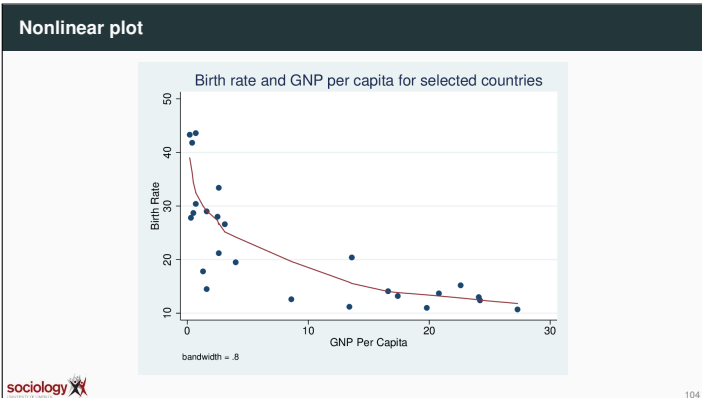
Birth rate and GNP example

```

do http://teaching.sociology.ul.ie/so5032/birth
sort gnp
label var bir "Birth Rate"
label var gnp "GNP Per Capita"
lowess bir gnp, title("Birth rate and GNP per capita for selected countries")

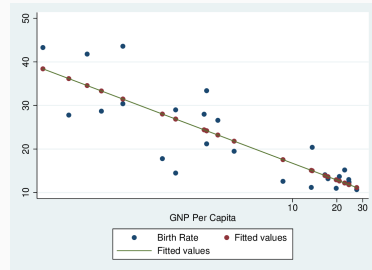
```

sociology
103



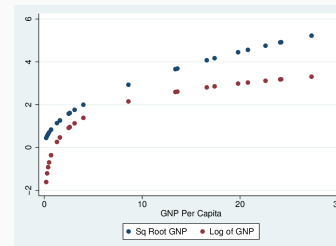
## Log-scale plot

```
scatter bir plog gnp, xscale(log) || line plog gnp, xscale(log)
```



## Square root and log compared

```
label var sqg "Sq Root GNP"
label var lg "Log of GNP"
scatter sqg lg gnp
```



## Residuals

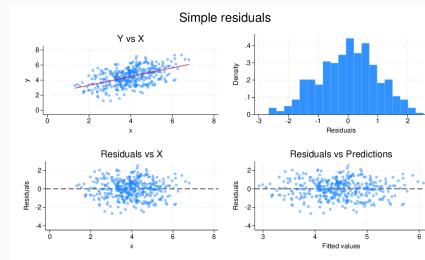
$$Y = b_0 + b_1 X_1 + \dots + b_k X_k + e$$

$$e \sim N(0, \sigma)$$

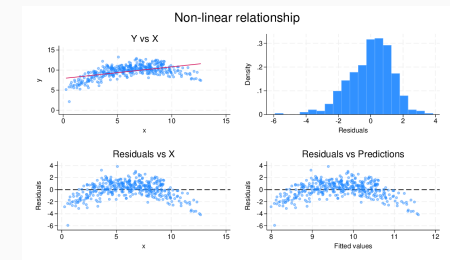
## Characteristics

- Residuals will
  - have mean 0
  - be as small as possible
  - have no linear relationship to X variables
- Residuals should
  - be approximately normally distributed (symmetric is often enough)
  - not have a non-linear relationship to any X variable
  - have a constant spread, that is not related to X or Y values
- If correlated with variables not in the model, perhaps those variables should be included

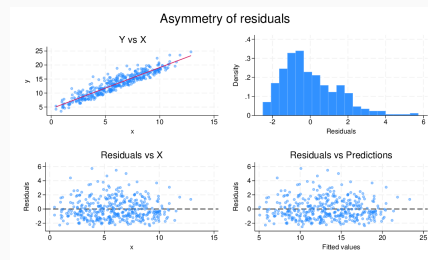
## Examining residuals: ideal



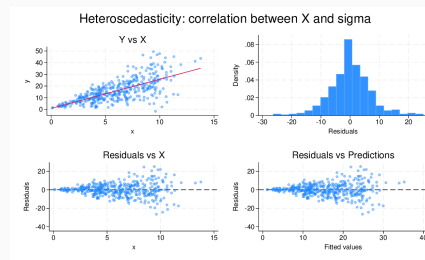
## Examining residuals: Non-linear



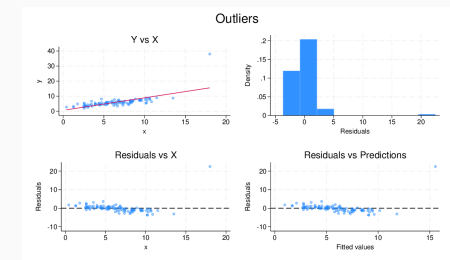
## Examining residuals: asymmetric



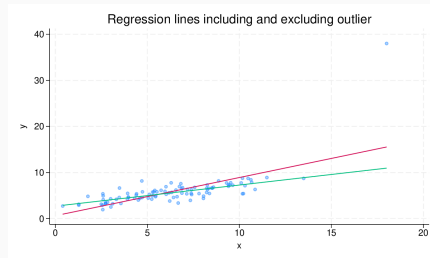
## Examining residuals: heteroscedasticity



## Examining residuals: Spotting outliers



## Examining residuals: Influence of outliers



## Lecture 6: Residuals and Influence

### Influence

## Outliers may have undue influence

- dfbeta
- Cook's distance

## DFBETA

- For each variable in the regression, for each case
- The effect of dropping that case on that variable
- Scaled by the standard error:

$$\frac{b - b^*}{SE}$$

## Cook's Distance

- A single number summarising each case's overall influence
- A scaled sum of changes in predicted Y

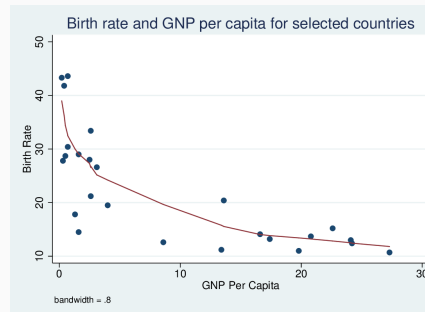
## Outlier interactive app

<https://teaching.sociology.ul.ie/apps/influence/>

## Birth rate and GNP example

```
do http://teaching.sociology.ul.ie/so5032/birth
sort gnp
label var bir "Birth Rate"
label var gnp "GNP Per Capita"
lowess bir gnp, title("Birth rate and GNP per capita for selected countries")
```

## Nonlinear plot



## Get linear relationship

```
reg bir gnp
```

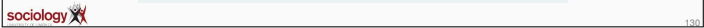
Source	SS	df	MS	Number of obs	=	25
Model	1450.2893	1	1450.2893	F(1, 23)	=	27.52
Residual	1212.02523	23	52.696749	Prob > F	=	0.0000
				R-squared	=	0.5447
				Adj R-squared	=	0.5249
Total	2662.28552	24	110.928563	Root MSE	=	7.2593

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
bir					
gnp	-.8139082	.155033	-5.25	0.000	-1.134018 - .4925981
_cons	29.6227	2.037416	14.54	0.000	25.40798 33.83742

```
predict plin
scatter bir plin gnp|| line plin gnp
```

## Linear plot



## Quadratic

Linear regression doesn't fit well

Clearly, as GNP rises BIR falls, but the rate of fall declines

Let's try quadratic:


```
reg bir c.gnp##c.gnp
```

```
. reg bir c.gnp##c.gnp
```

Source	SS	df	MS	Number of obs	=	25
Model	1665.82856	2	832.914278	F(2, 22)	=	18.39
Residual	996.456968	22	45.2934985	Prob > F	=	0.0000
				R-squared	=	0.6257
				Adj R-squared	=	0.5917
Total	2662.28552	24	110.928663	Root MSE	=	6.73

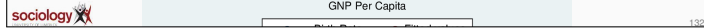
  

	bir	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
	gnp	-2.130192	.6205087	-3.43	0.002	-3.417048 - .8433351


131

## Quatratric plot

```
predict pquad
scatter bir pquad gnp|| line pquad gnp
```



$\sqrt{GNP}$

Let's try square root of GNP:

```
gen sqg = sqrt(gnp)
reg bir sqg
```

```


. gen sqg = sqrt(gnp)
. reg bir sqg

```

Source	SS	df	MS	Number of obs	=	25
				F(1, 23)	=	39.44
Model	1681.66084	1	1681.66084	Prob > F	=	0.0000
Residual	980.62466	23	42.638569	R-squared	=	0.6317
				Adj R-squared	=	0.6156
Total	2662.28552	24	110.928563	Root MSE	=	6.5296

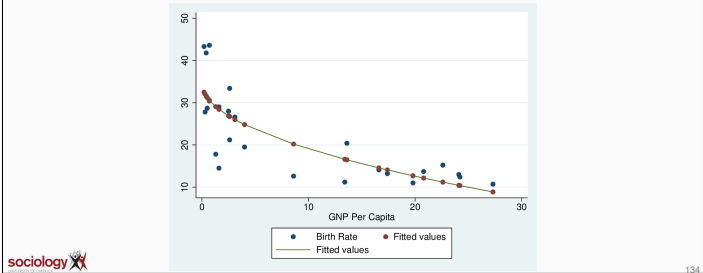
  

	bir	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
sqg		-4.945487	.7874579	-6.28	0.000	-6.574468 -3.316506
_cons		34.70314	2.391073	14.51	0.000	29.75683 39.64946

**sociology** 

$\sqrt{GNP}$  plot

```
predict psqrt
scatter bir psqrt gnp|| line psqrt gnp
```



## log(GNP)

Let's try the log of GNP:

```
gen lgg = log(gnp)
reg bir lgg
```

```


. gen lgg = log(gnp)
. reg bir lgg

```

Source	SS	df	MS	Number of obs	=	25
Model	1875.68482	1	1875.68482	F(1, 23)	=	54.84
Residual	786.600705	23	34.200307	Prob > F	=	0.0000
				R-squared	=	0.7045
				Adj R-squared	=	0.6917
Total	2662.28552	24	110.928563	Root MSE	=	5.8481

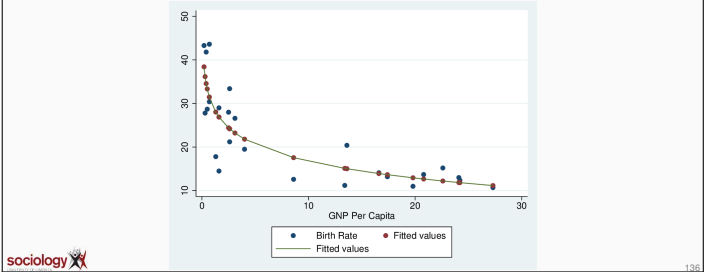
	bir	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lgg		-5.542152	.748362	-7.41	0.000	-7.090257 -3.994047
_cons		29.49466	1.53576	19.21	0.000	26.3177 32.67162


Downloaded from <https://www.cambridge.org/core>

1/35

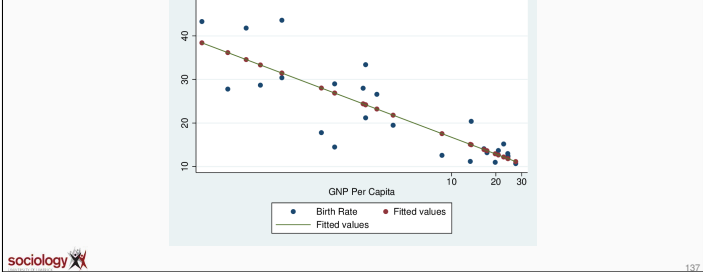
## log(GNP) plot

```
predict plog
scatter bir plog gnp|| line plog gnp
```



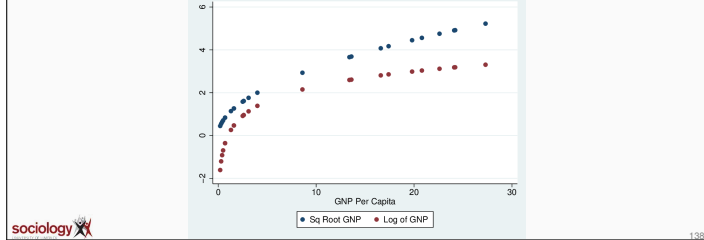
## Log-scale plot

```
scatter bir plog gnp, xscale(log) || line plog gnp, xscale(log)
```



## Square root and log compared

```
label var sqg "Sq Root GNP"
label var lg "Log of GNP"
scatter sqg lg gnp
```





## Lecture 7: Logs and log regression

### Logarithms

#### Logarithms

Logarithms allow us to move between multiplicative equations and additive ones.

Logs are defined relative to a base number. If we take 10 as the base then  $y = \log_{10}(x)$  means  $10^x = y$ .

It's easy to calculate the log of powers of 10:

$$\begin{array}{ll} \log(10) = 1 & 10^1 = 10 \\ \log(100) = 2 & 10^2 = 100 \\ \log(1000) = 3 & 10^3 = 1000 \\ \log(1000000) = 6 & 10^6 = 1000000 \end{array}$$

$10^0$  is defined as 1, so the log of 1 is zero.

#### From 0 to 1

For numbers between 1 and 0, logs are negative

$$\begin{array}{ll} \frac{1}{10} = 10^{-1} & \log(0.1) = -1 \\ \frac{1}{100} = 10^{-2} & \log(0.01) = -2 \\ \frac{1}{1000} = 10^{-3} & \log(0.001) = -3 \end{array}$$

The  $\log_{10}$  of powers of 10 are integers, but we can raise 10 to non-integer powers too, to get the log of any number greater than zero. For instance,  $10^{2.09}$  is 123, so the log of 123 is 2.09.

#### Multiply by adding

We can see with round powers of 10 than using logs we can move between multiplication and addition:

$$\begin{array}{l} 100 \times 1000 = 100000 \\ 10^2 \times 10^3 = 10^5 = 10^{2+3} \end{array}$$

#### Calculate A × B

Thus do calculate A × B we do as follows:

- Calculate  $\log(A)$
- Calculate  $\log(B)$
- Calculate  $\log(C) = \log(A) + \log(B)$
- Take the anti-log of  $\log(C)$ , i.e.,  $10^{\log(C)} = C$

#### Example

Multiply 12345 by 67890

$$\log(12345) = 9.421$$

$$\log(67890) = 11.126$$

$$9.421 + 11.126 = 20.547$$

$$10^{20.547} = 838102050$$

#### An application

If you have a certain quantity (e.g., money in a bank account), whose value increases by a constant proportion every year, its value in any year depends on a multiplicative relationship.

Let's say the increases is  $\alpha$  (i.e., a 10% increase means  $\alpha = 1.1$ )

#### Compound interest

$$\begin{array}{ll} \text{Year 0} & 100 \\ \text{Year 1} & 100 \times \alpha \\ \text{Year 2} & 100 \times \alpha \times \alpha \\ \text{Year 3} & 100 \times \alpha \times \alpha \times \alpha \\ \text{Year 4} & 100 \times \alpha \times \alpha \times \alpha \times \alpha \\ \text{Year 5} & 100 \times \alpha \times \alpha \times \alpha \times \alpha \times \alpha \end{array}$$

In short, the value in year  $t$  is  $100 \times \alpha^t$

$$y_t = 100 \times \alpha^t$$

#### Constant proportional increase

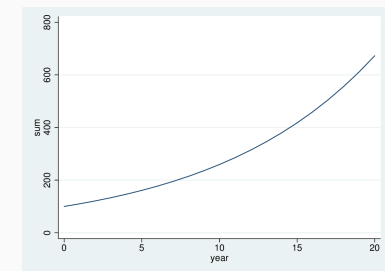


Figure 1: A constant proportional increase

## Convert to logs

But if we convert to logs we can calculate it as follows

$$\log(y_t) = \log(100) + t \times \log(\alpha)$$

In other words, rather than multiplying by  $\alpha$  every year, we add  $\log(\alpha)$ .

## Plot

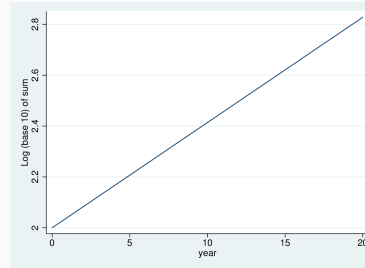


Figure 2: Taking the base-10 log of the sum: a straight line

## Straight line

This gives a straight line relationship (see Fig 2).

Thus we can use logs to move between multiplicative and additive (straight-line) relationships.

## Other bases

Logs to the base 10 are easy to understand, but the base number need not be 10. A log to the base  $n$  is defined thus:

$$y = \log_n(x) \Leftrightarrow n^y = x$$

## Natural logs

Computer scientists often use  $\log_2$ , but the most common log base is the special number  $e \approx 2.7183$ . This has some special mathematical properties that make certain calculations easier.

Logs to base  $e$  are called natural logs, often written  $\ln(x)$  etc:

$$y = \ln(x) \Leftrightarrow e^y = x$$

See Fig 3, which shows that the natural log also gives a straight line.

## Natural log straight line

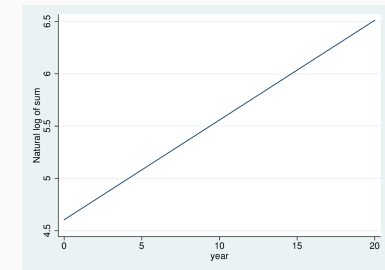


Figure 3: Taking the natural log of the sum: also a straight line

## Natural log

- Fig 4 shows the natural log of  $X$  from 0.1 (-2.303) to 100 (4.605).
- For  $X = 1$ , the log is 0.
- As  $X$  approaches 0, the log falls faster and faster.
- As  $X$  rises above 1, the log rises, but more slowly as it goes.
- Note that the log rises from  $X = 5$  to 10 as much as it does from  $X = 40$  to 80.

## $X$ vs $\ln(X)$

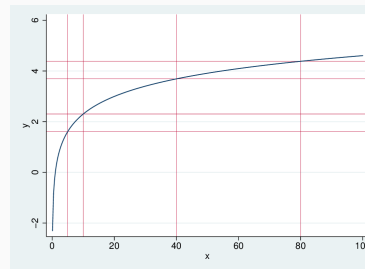


Figure 4: The natural log of  $X$  for  $X$  from 0.1 to 100

## Lecture 7: Logs and log regression

### Early pandemic: exponential curves

## Logs and COVID-19

- In the early stage of an epidemic, infections tend to increase at a steady rate
- On average each infected person infects others at a given rate, e.g., one person every four days
- So numbers of cases tend to rise at a steady percentage
  - New infections are proportional to existing infections
  - 100 today means 125 tomorrow, 156 the next day, etc.

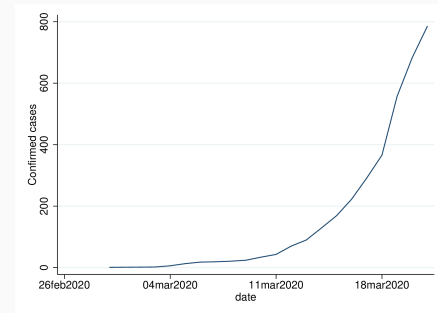
## Confirmed cases in Ireland

If we look at the raw number of cases in Ireland:

- it starts off very low
- stays there for a while
- but then starts rising
- and rising faster and faster

```
line cases date
```

## Confirmed cases in Ireland



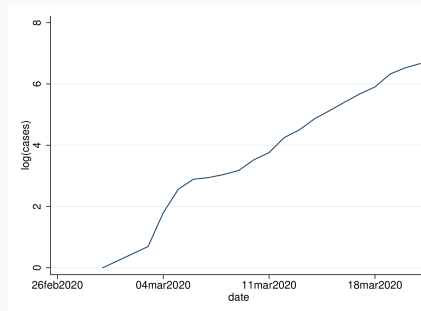
## Log cases

If we plot the log of the cases we see a different picture

- wobbly to begin with
- then approximating a straight line

```
gen lcases = log(cases)
line lcases date
```

## Log cases



## Log cases: straight => exponential

A straight line in logs means  $\log(ncases)$  increases by more or less a set amount every day

That means  $ncases$  rises by a set proportion every day: exponential rise

Exponential: even if it starts small, if given long enough, will get very very big!

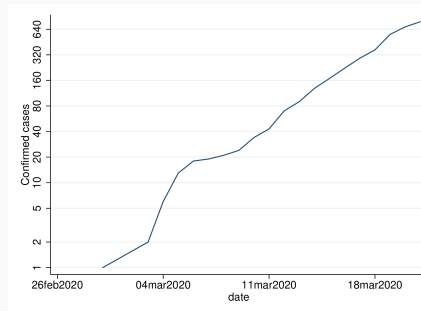
## Log scale, real cases

We can graph  $\log(cases)$  but we can also graph cases with a Y log-scale

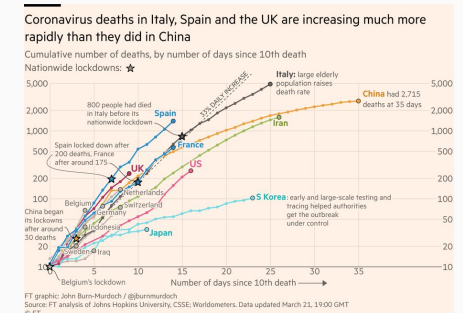
```
line cases date, yscale(log) ylabel(1 2 5 10 20 40 80 160 320 640)
```

This gives the advantages of the logging while retaining the real numbers on the axis

## Log scale, real cases



## Log-scale graphic in the wild



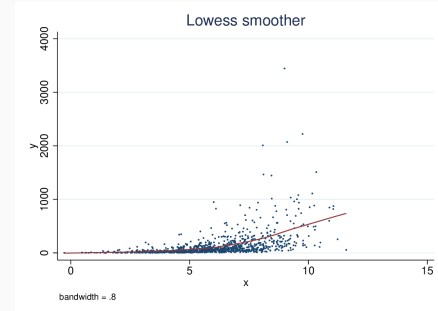
## Lecture 7: Logs and log regression

### Log regression

### Multiplicative relationship

- Where the underlying relationship is multiplicative, linear regression doesn't work well
- Implies an additive increase where a multiplicative one is better
- If we take the log of the dependent variable:
  - better estimates
  - often cures heteroscedasticity

### Simulation: Y increases 65% for X +1



### Linear regression

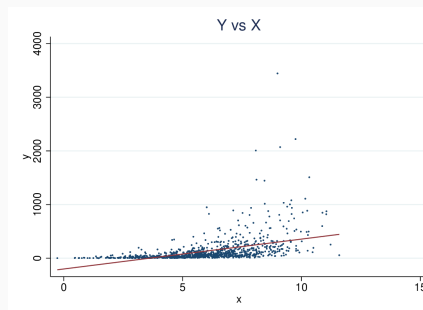
```
. reg y x
```

Source	SS	df	MS	Number of obs	=	1,000
Model	12181477.5	1	12181477.5	F(1, 998)	=	274.71
Residual	44253675.2	998	44342.3599	Prob > F	=	0.0000
				R-squared	=	0.2158
				Adj R-squared	=	0.2151
Total	56435152.7	999	56491.6443	Root MSE	=	210.58

	y	Coefficient	Std. err.	t	P> t	[95% conf. interval]
x		55.69088	3.360033	16.57	0.000	49.09734 62.28442
_cons		-200.7041	20.95566	-9.58	0.000	-241.8263 -159.5819

### Predictions



### Log(Y)

```
. gen ly = log(y)
. reg ly x
```

Source	SS	df	MS	Number of obs	=	1,000
Model	956.12538	1	956.12538	F(1, 998)	=	1032.66
Residual	924.030142	998	.925881905	Prob > F	=	0.0000
				R-squared	=	0.5085
				Adj R-squared	=	0.5080
Total	1880.15552	999	1.88203756	Root MSE	=	.96223

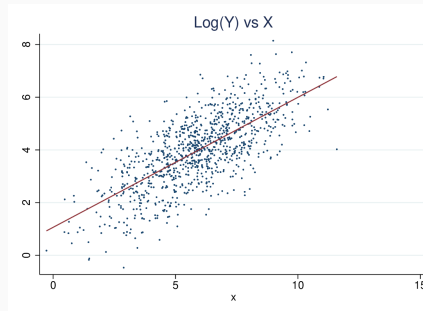
  

	ly	Coefficient	Std. err.	t	P> t	[95% conf. interval]
x		.4933914	.0153537	32.14	0.000	.4632622 .5235205
_cons		1.062305	.0957568	11.09	0.000	.8743972 1.250213

### Interpretation

- For a 1 unit change in X,  $\log(\hat{Y})$  rises by 0.4933914
- Thus for a 1 unit change in X, Y rises by  $e^{0.4933914} = 1.638$
- $e^{0.4933914}$  is the antilog of 0.4933914

### Predictions



### Predicted values

- Where the dependent variable is logged the prediction of the Y value is not simply the anti-log of the predicted  $\log(Y)$
- When we take the anti-log we must take account of the fact that residuals above the line expand by more than residuals below the line
- Thus a small correction

$$\log(\hat{Y}) = a + bX$$

$$\hat{Y} = e^{\log(\hat{Y})} * e^{RMSE^2/2}$$

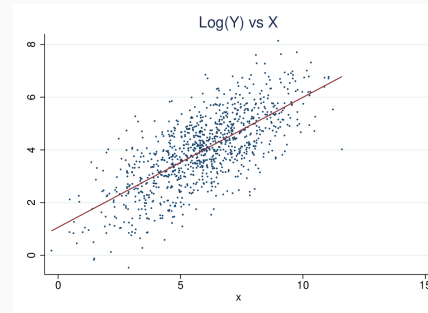
- where RMSE is the standard deviation of the regression

## Calculations

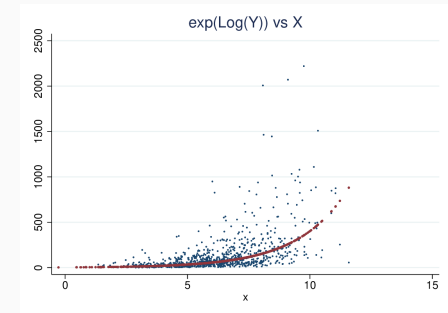
```
gen ly = log(y)
reg ly x

predict lyhat
gen elyh = exp(lyhat)
gen elyh2 = elyh * exp(rmse^2/2)
```

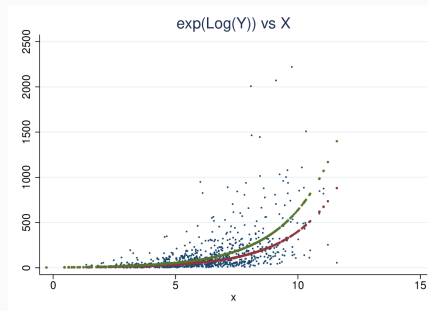
## Predictions: predict log(Y) on log scale



## Predictions: only $e^{\log(Y)}$



## Predictions: with correction



## Predicting COVID-19

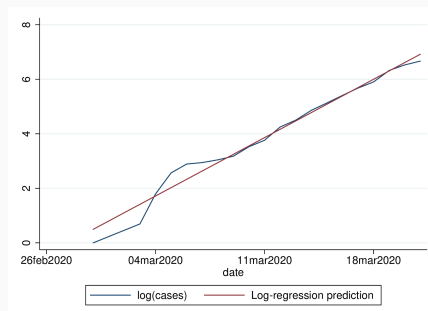
- We can apply log regression to the COVID-19 data
- A straight line on a log scale means a constant proportional increase.
- We can estimate this increase, regressing  $\log(\text{cases})$  on date.
- The slope,  $b$ , is the amount by which  $\log(\hat{\text{cases}})$  rises per day
- $e^b$  is then the multiplier by which cases rises per day

```
reg lcases date
```

## Stata output

. reg lc date					
Source	SS	df	MS	Number of obs =	20
Model	66.1088015	1	66.1088015	F(1, 18) =	746.82
Residual	1.59336573	18	.088520318	P>F =	0.0000
Total	67.7021673	19	3.56327196	R-squared =	0.9765
				Adj R-squared =	0.9752
				Root MSE =	.29752
lc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
date	-.3058309	.0111911	-27.33	0.000	-.2823193 - .3293426
_cons	-6719.833	246.0411	-27.31	0.000	-7236.746 - 6202.92

## Logs with log regression



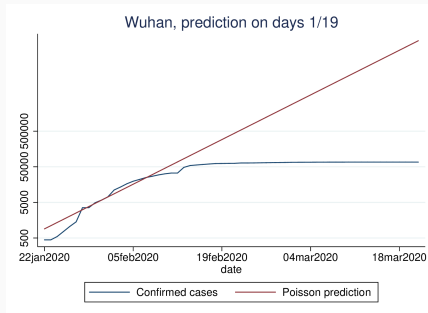
## Steady increase

The log of cases rises by 0.3058 per day  
 This means cases rises by a factor of  $e^{0.3058} = 1.358$   
 The increase is  $1.358 - 1 = 0.358$ , or almost 36% per day  
 Implies a doubling about every 2.6 days

## But exponential increase is temporary

Exponential increase cannot go on indefinitely  
 Even if nothing is done, the rate of increase will decline as fewer people are left unexposed  
 And interventions (isolation, tracing) will reduce the rate  
 See China, for example

## Wuhan, with prediction based on 1st 19 days



## Summary

If there is a constant rate of increase, logs give us straight lines  
 Graph the log, or use a log scale on the Y-axis  
 Log regression allows us to estimate the rate  
 Exponential increase isn't forever, but modelling the exponential helps us see where the rate starts to drop  
 Code available here: <http://teaching.sociology.ul.ie/so5032/irecovid.do>

## Outline

Today we introduce logistic regression: for binary outcomes  
 See Agresti Ch 15 Sec 1.

## Binary outcomes and regression

- OLS (linear regression) requires an interval dependent variable
- Binary or "yes/no" dependent variables are not suitable
- Nor are rates, e.g., n successes out of m trials

## Problems with OLS

- Errors are distinctly not normal
- While predicted value can be read as a probability, can depart from 0:1 range
- Particular difficulties with multiple explanatory variables
- Nonetheless still often used

## Linear Probability Model

- If we use OLS with binary outcomes, it is called "linear probability model":

$$Pr(Y = 1) = a + bX$$

- data is 0/1, prediction is probability
- Assumptions violated, but if predicted probabilities in range 0.2–0.8, not too bad

## Credit card example

```

. reg card income

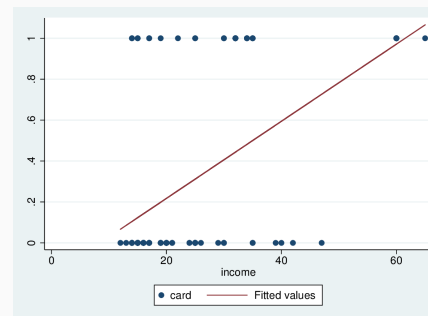
```

Source	SS	df	MS	Number of obs =	100
Model	5.55556122	1	5.55556122	F(1, 98)	= 34.38
Residual	15.0344388	98	.161575906	Prob > F	= 0.0000
Total	21.39	99	.216060606	R-squared	= 0.2597
				Adj R-squared	= 0.2522
				Root MSE	= .40197

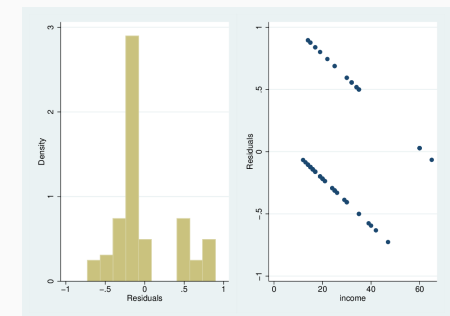
  

	card	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
income		.0188458	.003214	5.86	0.000	.0124678 .0252238
_cons		-.1594495	.089584	-1.78	0.078	-.3372261 .018327

## Credit card example



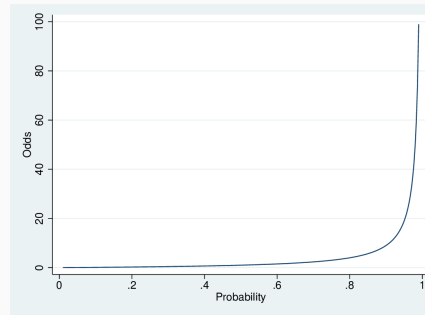
## Credit card example



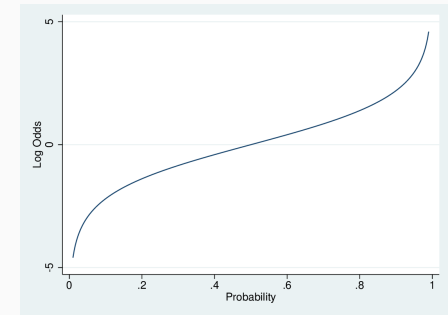
## Logistic transformation

- Probability is bounded  $[0 : 1]$
- OLS predicted value is unbounded
- How to transform probability to  $-\infty : \infty$  range?
- Odds:  $\frac{p}{1-p}$  – range is  $0 : \infty$
- Log of odds:  $\log \frac{p}{1-p}$  has range  $-\infty : \infty$

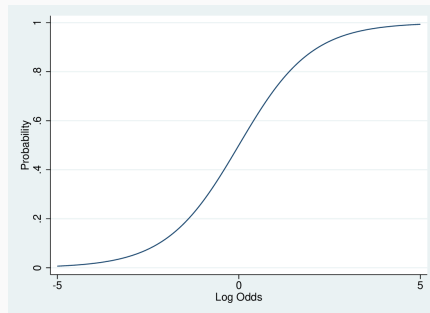
## Probability to odds



## Probability to log-odds



## Rotated: the "S-shaped" curve



## Logistic regression

- Logistic regression uses this as the dependent variable:

$$\log \left( \frac{p}{1-p} \right) = a + bX$$

## Alternatives

We can look at this in three ways

- In terms of log-odds:

$$\log \left( \frac{Pr(Y=1)}{1-Pr(Y=1)} \right) = a + bX$$

- In terms of odds:

$$\frac{Pr(Y=1)}{1-Pr(Y=1)} = e^{a+bX}$$

- In terms of probability:

$$Pr(Y=1) = \frac{e^{a+bX}}{1 + e^{a+bX}} = \frac{1}{1 + e^{-a-bX}}$$

## Parameters

- The b parameter is the effect of a unit change in X on  $\log \left( \frac{Pr(Y=1)}{1-Pr(Y=1)} \right)$
- This implies a multiplicative change of  $e^b$  in  $\frac{Pr(Y=1)}{1-Pr(Y=1)}$ , in the Odds
- Thus an odds ratio
- But the effect of b on P depends on the level of b

## Credit card logistic regression

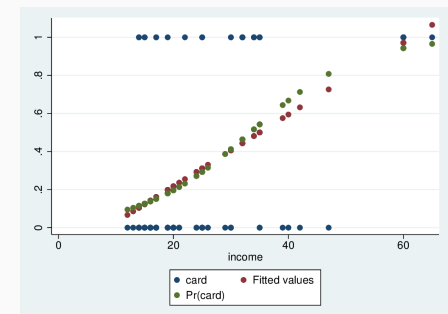
```
. logit card income
Iteration 0: log likelihood = -61.910066
Iteration 1: log likelihood = -48.707265
Iteration 2: log likelihood = -48.613215
Iteration 3: log likelihood = -48.61304
Iteration 4: log likelihood = -48.61304

Logistic regression              Number of obs   =      100
                                LR chi2(1)          =      26.59
                                Prob > chi2         =      0.0000
                                Pseudo R2          =      0.2148

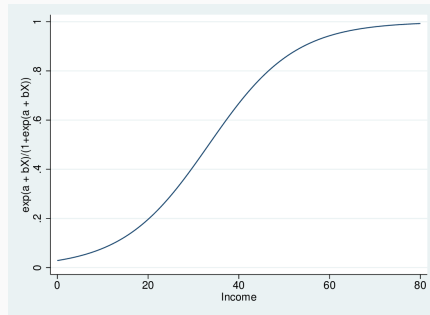
Log likelihood = -48.61304
```

card	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
income	.1054089	.0261574	4.03	0.000	.0541413 .1566765
_cons	-3.517947	.7103358	-4.95	0.000	-4.910179 -2.125714

## Credit card logistic regression



## Sigmoid curve from $a+bX$



## Calculating predicted probabilities by hand

- We can calculate the predicted probability for any combination of values of the independent variables
- First, plug them into the  $a + bX$  part to get the predicted log-odds
- Then take the anti-log of the log-odds to get the odds
- Then  $\text{odds}/(1+\text{odds})$  gives us the probability

## Calculating predicted probabilities

- Example:  $\log(\text{odds}) = 0.25 + 0.12X$
- Predict for  $X = 10$ 
  - Predicted log-odds =  $0.25 + 0.12 \cdot 10 = 1.45$
  - Predicted odds =  $e^{1.45} = 4.263$
  - Predicted probability =  $4.263 / (1 + 4.263) = 0.810$

## Web applet for practicing

<https://teaching.sociology.ul.ie/apps/logabx/>

## Outline

Today we introduce logistic regression: for binary outcomes  
See Agresti Ch 15 Sec 1.

## Binary outcomes and regression

- OLS (linear regression) requires an interval dependent variable
- Binary or “yes/no” dependent variables are not suitable
- Nor are rates, e.g.,  $n$  successes out of  $m$  trials

## Problems with OLS

- Errors are distinctly not normal
- While predicted value can be read as a probability, can depart from 0:1 range
- Particular difficulties with multiple explanatory variables
- Nonetheless still often used

## Linear Probability Model

- If we use OLS with binary outcomes, it is called “linear probability model”:

$$Pr(Y = 1) = a + bX$$

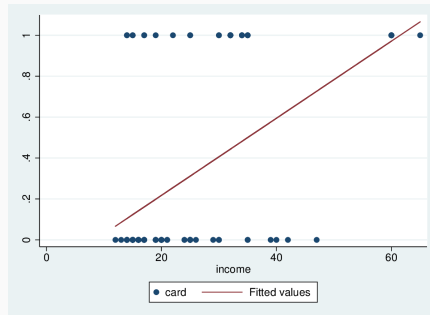
- data is 0/1, prediction is probability
- Assumptions violated, but if predicted probabilities in range 0.2–0.8, not too bad

## Credit card example

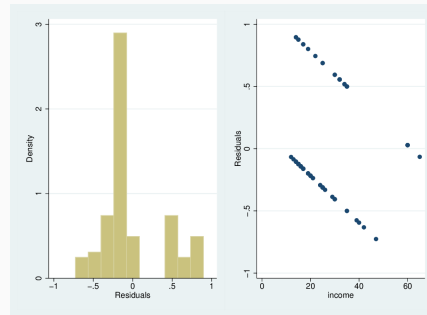
. reg card income					
Source	SS	df	MS	Number of obs	= 100
Model	5.55556122	1	5.55556122	F(1, 98)	= 94.38
Residual	15.8944388	98	.161575906	Prob > F	= 0.0000
Total	21.39	99	.216060606	R-squared	= 0.2597
				Adj R-squared	= 0.2522
				Root MSE	= .40197
card	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
income	.0188458	.003214	5.86	0.000	.0124678 .0252238
_cons	-.1594495	.089584	-1.78	0.078	-.3372261 .018327



## Credit card example



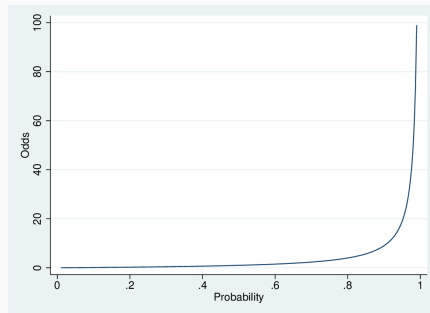
## Credit card example



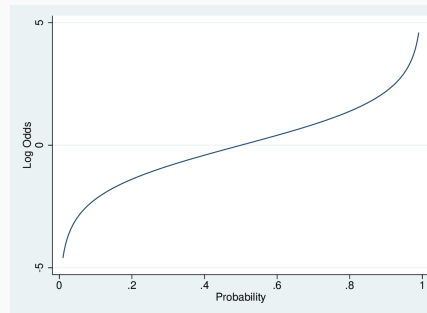
## Logistic transformation

- Probability is bounded  $[0 : 1]$
- OLS predicted value is unbounded
- How to transform probability to  $-\infty : \infty$  range?
- Odds:  $\frac{p}{1-p}$  - range is  $0 : \infty$
- Log of odds:  $\log \frac{p}{1-p}$  has range  $-\infty : \infty$

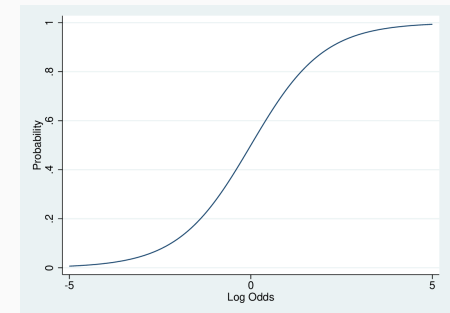
## Probability to odds



## Probability to log-odds



## Rotated: the "S-shaped" curve



## Logistic regression

- Logistic regression uses this as the dependent variable:

$$\log \left( \frac{p}{1-p} \right) = a + bX$$

## Alternatives

We can look at this in three ways

- In terms of log-odds:

$$\log \left( \frac{Pr(Y=1)}{1-Pr(Y=1)} \right) = a + bX$$

- In terms of odds:

$$\frac{Pr(Y=1)}{1-Pr(Y=1)} = e^{a+bX}$$

- In terms of probability:

$$Pr(Y=1) = \frac{e^{a+bX}}{1 + e^{a+bX}} = \frac{1}{1 + e^{-a-bX}}$$

## Parameters

- The b parameter is the effect of a unit change in X on  $\log \left( \frac{Pr(Y=1)}{1-Pr(Y=1)} \right)$
- This implies a multiplicative change of  $e^b$  in  $\frac{Pr(Y=1)}{1-Pr(Y=1)}$ , in the Odds
- Thus an odds ratio
- But the effect of b on P depends on the level of b

## Credit card logistic regression

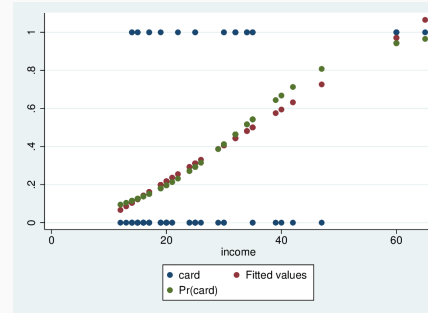
```
. logit card income
Iteration 0: log likelihood = -61.910066
Iteration 1: log likelihood = -48.707265
Iteration 2: log likelihood = -48.613215
Iteration 3: log likelihood = -48.61304
Iteration 4: log likelihood = -48.61304

Logistic regression              Number of obs   =      100
                                LR chi2(1)        =      26.59
                                Prob > chi2       =      0.0000
                                Pseudo R2        =      0.2148

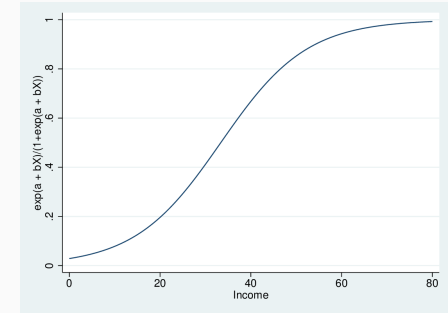
Log likelihood = -48.61304
```

	card	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
	income	.1054089	.0261574	4.03	0.000	.0541413 .1566765
	_cons	-3.517947	.7103358	-4.95	0.000	-4.910179 -2.125714

## Credit card logistic regression



## Sigmoid curve from a+bX



## Calculating predicted probabilities by hand

- We can calculate the predicted probability for any combination of values of the independent variables
- First, plug them into the  $a + bX$  part to get the predicted log-odds
- Then take the anti-log of the log-odds to get the odds
- Then  $\text{odds} / (1 + \text{odds})$  gives us the probability

## Calculating predicted probabilities

- Example:  $\log(\text{odds}) = 0.25 + 0.12X$
- Predict for  $X = 10$ 
  - Predicted log-odds =  $0.25 + 0.12 \cdot 10 = 1.45$
  - Predicted odds =  $e^{1.45} = 4.263$
  - Predicted probability =  $4.263 / (1 + 4.263) = 0.810$

## Web applet for practicing

<http://teaching.sociology.ul.ie:3838/logabx/>

## Housing tenure

- Housing tenure: probability of owning outright, BHPS data

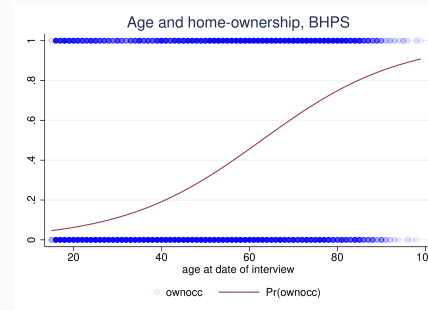
```
. logit ownocc age
Iteration 0: Log likelihood = -8728.6773
Iteration 1: Log likelihood = -7150.2389
Iteration 2: Log likelihood = -7095.7194
Iteration 3: Log likelihood = -7095.5268
Iteration 4: Log likelihood = -7095.5268

Logistic regression              Number of obs   = 14,182
                                LR chi2(1)        = 3266.30
                                Prob > chi2       = 0.0000
                                Pseudo R2        = 0.1871

Log likelihood = -7095.5268
```

	ownocc	Coefficient	Std. err.	z	P> z	[95% conf. interval]
	age	.0633183	.0012705	49.84	0.000	.0608281 .0658094
	_cons	-3.974023	.0697795	-56.95	0.000	-4.110788 -3.837258

## Predictions



## Predictions

$$LO = a + bX$$

$$\text{Odds} = \exp(a + bX)$$

$$P = \text{Odds} / (1 + \text{Odds})$$

X increases by 1:

- LO by b (additive)
- Odds by  $e^b$  (multiplicative)
- P is more complicated

Prediction

- Log-odds

$$X = x \quad LO(x) = a + bx$$

$$X = x+1 \quad LO(x+1) = a + b(x+1) = a + bx + b$$

$$\text{Difference: } LO(x+1) - LO(x) = b$$

sociology

226

Prediction: odds scale

- Odds

$$X = x \quad Odds(x) = e^{a+bx} = e^a e^{bx}$$

$$X = x+1 \quad Odds(x+1) = e^{a+b(x+1)} = e^{a+bx+b} = e^a e^{bx} e^b$$

$$\text{Ratio} \quad Odds(x+1)/Odds(x) = e^b$$

- Hence odds-ratio: if X increases by 1, OR increases by factor of  $e^b$

sociology

227

Odds ratio

```
. use univ, replace
. logit ownocc i.univ
Iteration 0: Log likelihood = -8729.869
Iteration 1: Log likelihood = -8710.8025
Iteration 2: Log likelihood = -8710.8469
Iteration 3: Log likelihood = -8710.8469
Logistic regression
Number of obs = 14,183
LR chi2(1) = 38.93
Prob > chi2 = 0.0000
Pseudo R2 = 0.0022
```

	univ	ownocc	z	P> z	[95% conf. interval]		
1.univ		.3336103	.0951897	-6.06	0.000	-4.417693	-.2264922
_cons		-.7702941	-.019524	-39.78	0.000	-8.149376	-.7380566

$$OR = (499/1514) / (3835/8335) = 0.7163$$

$$e^b = e^{-.3336103} = 0.7163$$

sociology

228

Predictions on probability scale

- Effect of X on the probability scale is non-linear
- Low when p is either high or low
- Highest at p = 0.5, odds = 1, log-odds = 0
- The steepest slope is at p = 0.5, with a value of  $\frac{\partial}{4}$

sociology

229

Marginal effects

Marginal effects of age on probability at 25, 62.76 and 90

sociology

230

Multiple explanatory variables

```
. logit ownocc age i.univ
Iteration 0: Log likelihood = -8728.6773
Iteration 1: Log likelihood = -7150.3435
Iteration 2: Log likelihood = -7094.4048
Iteration 3: Log likelihood = -7094.1883
Iteration 4: Log likelihood = -7094.1882
Logistic regression
Number of obs = 14,182
LR chi2(2) = 3268.98
Prob > chi2 = 0.0000
Pseudo R2 = 0.1873
```

	ownocc	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
age		.0636471	.0012888	49.38	0.000	.061121	.0661731
1.univ		.0999785	.0608614	1.64	0.100	-.0193076	.2192646
_cons		-4.004807	.0724889	-55.25	0.000	-4.146883	-3.862731

sociology

231

Lecture 11: Multinomial and Ordinal regression

Inference

sociology

232

Inference

- In practice, inference is similar to OLS though based on a different logic
- For each explanatory variable,  $H_0 : \beta = 0$  is the interesting null
- $z = \frac{\hat{\beta}}{SE}$  is approximately normally distributed (large sample property)
- More usually, the Wald test is used:  $\left(\frac{\hat{\beta}}{SE}\right)^2$  has a  $\chi^2$  distribution with one degree of freedom

sociology

232

Likelihood ratio tests

- The "likelihood ratio" test is thought more robust than the Wald test for smaller samples
- Where  $l_0$  is the likelihood of the model without  $X_j$ , and  $l_1$  that with it, the quantity
$$-2 \left( \log \frac{l_0}{l_1} \right) = -2 (\log l_0 - \log l_1)$$
is  $\chi^2$  distributed with one degree of freedom

sociology

233

## Nested models

- More generally,  $-2 \left( \log \frac{L_1}{L_0} \right)$  tests nested models: where model 1 contains all the variables in model 0, plus  $m$  extra ones, it tests the null that all the extra  $\beta$  coefficients are zero ( $\chi^2$  with  $m$  df)
- If we compare a model against the null model (no explanatory variables, it tests

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_K = 0$$

- Strong analogy with  $F$  test in OLS

## Example

```
. qui logit owncc age
. est store m0d1
. logit owncc age i.educ
Iteration 0: Log likelihood = -8758.8773
Iteration 1: Log likelihood = -7138.2084
Iteration 2: Log likelihood = -7077.7732
Iteration 3: Log likelihood = -7077.8203
Iteration 4: Log likelihood = -7077.8203

Logistic regression              Number of obs = 14,182
                                LR chi2(3) = 3302.31
                                Prob > chi2 = 0.0000
                                Pseudo R2 = 0.1892

Log likelihood = -7077.8203

+-----+-----+
| owncc | Coefficient | Std. err. | z  | P>|z| | [95% conf. interval] |
+-----+-----+
| age   | .0655989   | .0151433   | 4.34 | 0.000 | .0359271   .0952707 |
+-----+-----+
| educ  |           |           |     |       |               |
| 1     | -.3041899  | .0733504   | -4.15 | 0.000 | -.4481642  -.1602156 |
| 2     | -.1076582  | .0461399   | -2.33 | 0.020 | -.2000077  -.0153087 |
| 3     | -.060816   | .0733524   | -0.83 | 0.408 | -0.203694  .0818604 |
+-----+-----+
| _cons | -4.060816  | .0733524   | -5.48 | 0.000 | -4.203694  -3.917933 |
+-----+-----+

. lrtest m0d1
Likelihood-ratio test
Assumption: m0d1 nested within .
Chi2(3) = 3302.31
Prob > chi2 = 0.0000
```

## Lecture 11: Multinomial and Ordinal regression

### Margins command

## "Average Marginal Effect"

- "What would happen to the average predicted probability if we increased X?"
- For linear regression, increase X by 1 => increase by  $b$ 
  - increase X by 10 => increase by  $b \times 10$
  - increase X by 0.1 => increase by  $b \times 0.1$
  - since it's a straight line
- For AME in logistic we use the slope of the tangent, for each X value
- Average across the observed data
- Gives something like a LPM slope

## AME in Stata

```
. margins, dydx(age)
Average marginal effects              Number of obs = 14,182
Model VCE: OIM
Expression: Pr(owncc), predict()
dy/dx wrt: age

+-----+-----+
|         | Delta-method |
| dy/dx   | std. err.   | z  | P>|z| | [95% conf. interval] |
+-----+-----+
| age     | .0104836    | .0001382 | 75.84 | 0.000 | .0102126   .0107545 |
+-----+-----+
```

## Lecture 11: Multinomial and Ordinal regression

### Maximum likelihood

## Maximum likelihood estimation

- What is this "likelihood"?
- Unlike OLS, logistic regression (and many, many other models) are estimated by *maximum likelihood estimation*
- In general this works by choosing values for the parameter estimates which maximise the probability (likelihood) of observing the actual data
- OLS can be ML estimated, and yields exactly the same results

## Iterative search

- Sometimes the values can be chosen analytically
  - A likelihood function is written, defining the probability of observing the actual data given parameter estimates
  - Differential calculus derives the values of the parameters that maximise the likelihood, for a given data set
- Often, such "closed form solutions" are not possible, and the values for the parameters are chosen by a systematic computerised search (multiple iterations)
- Extremely flexible, allows estimation of a vast range of complex models within a single framework

## Likelihood as a quantity

- Either way, a given model yields a specific maximum likelihood for a given data set
- This is a probability, hence bounded  $[0 : 1]$
- Reported as log-likelihood, hence bounded  $[-\infty : 0]$
- Thus is usually a large negative number
- Where an iterative solution is used, likelihood at each stage is usually reported – *normally* getting nearer 0 at each step

## Lecture 11: Multinomial and Ordinal regression

### Tabular data

### Tabular data

- If all the explanatory variables are categorical (or have few fixed values) your data set can be represented as a table
- If we think of it as a table where each cell contains  $n$  yeses and  $m - n$  noes ( $n$  successes out of  $m$  trials) we can fit grouped logistic regression
- $n$  successes out of  $m$  trials implies a binomial distribution of degree  $m$

$$\log \frac{n}{m-n} = \alpha + \beta X$$

- The parameter estimates will be exactly the same as if the data were treated individually

### Tabular data and goodness of fit

- But unlike with individual data, we can calculate goodness of fit, by relating observed successes to predicted in each cell
- If these are close we cannot reject the null hypothesis that the model is incorrect (i.e., you want a high p-value)
- Where  $l_i$  is the likelihood of the current model, and  $l_s$  is the likelihood of the "saturated model" the test statistic is

$$-2 \left( \log \frac{l_i}{l_s} \right)$$

- The saturated model predicts perfectly and has as many parameters as there are "settings" (cells in the table)
- The test has  $df$  of number of settings less number of parameters estimated, and is  $\chi^2$  distributed

### Grouped card data

```
. glm credit income, family(binomial n)
Iteration 0: Log likelihood = -27.687962
Iteration 1: Log likelihood = -27.416501
Iteration 2: Log likelihood = -27.416501
Iteration 3: Log likelihood = -27.416501

Generalized linear models
Optimization : ML
Number of obs = 24
Residual df = 22
Scale parameter = 1
Deviance = 39.275792 (1/df) Deviance = 1.785263
Pearson = 32.20690239 (1/df) Pearson = 1.468041
Variance function: V(u) = u*(1-u) [Binomial]
Link function : g(u) = ln(u/(1-u)) [Logit]
Log likelihood = -27.41650068
AIC = 2.451375
BIC = -30.64139
```

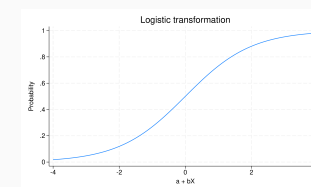
	GLM				
credit	Coefficient	std. err.	z	P> z	[95% conf. interval]
income	.1054089	.0261574	4.03	0.000	-.0541413 .1566765
_cons	-3.517947	.7103358	-4.96	0.000	-4.910179 -2.125714

## Lecture 11: Multinomial and Ordinal regression

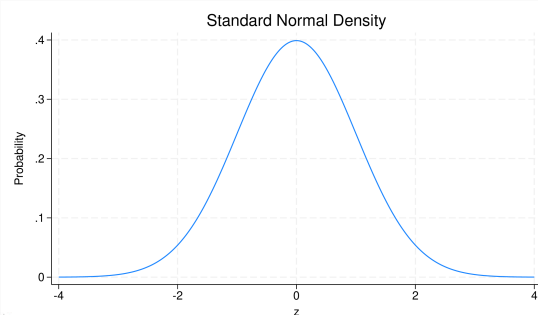
### Probit

### Alternatives to logistic: probit regression

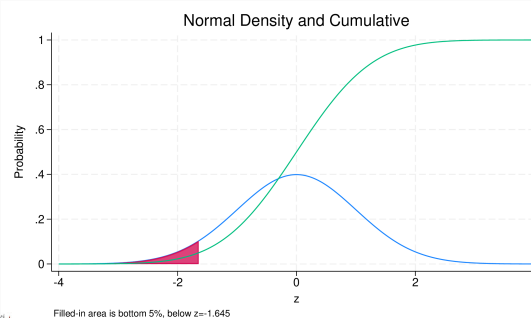
- The logistic transformation gives us an S-shaped curve relating  $a+bX$  to probability
- There are other ways of getting this curve



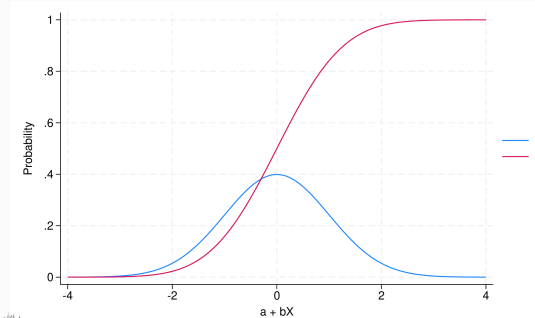
### The Standard Normal: density and CDF



### The Standard Normal: density and CDF



### Probit transformation



Card and income: logit

```

. logit card income
Iteration 0: Log likelihood = -61.910066
Iteration 1: Log likelihood = -48.707265
Iteration 2: Log likelihood = -48.613215
Iteration 3: Log likelihood = -48.61304
Iteration 4: Log likelihood = -48.61304

Logistic regression              Number of obs =   100
                                LR chi2(1)    =   26.59
                                Prob > chi2    =  0.0000
                                Pseudo R2     =  0.2148

Log likelihood = -48.61304

```

card	Coefficient	Std. err.	z	P> z	[95% conf. interval]
income	.1054089	.0261574	4.03	0.000	.0541413 .1566765
_cons	-3.517947	.7103358	-4.95	0.000	-4.910179 -2.125714

sociology

248

Card and income: probit

```

. probit card income
Iteration 0: Log likelihood = -61.910066
Iteration 1: Log likelihood = -48.59092
Iteration 2: Log likelihood = -48.550994
Iteration 3: Log likelihood = -48.550978
Iteration 4: Log likelihood = -48.550978

Probit regression              Number of obs =   100
                                LR chi2(1)    =   26.72
                                Prob > chi2    =  0.0000
                                Pseudo R2     =  0.2158

Log likelihood = -48.550978

```

card	Coefficient	Std. err.	z	P> z	[95% conf. interval]
income	.0622283	.0141879	4.39	0.000	.0344205 .0900361
_cons	-2.089336	.3821555	-5.47	0.000	-2.838347 -1.340326

sociology

249

Predictions

```

. di exp(-3.517947 + .1054089*50)/(1 + exp(-3.517947 + .1054089*50))
.8522676
. di normal(-2.089336 + .0622283*50)
.84662824

```

sociology

250

Card probability

sociology

251

Comparison

- Logit estimates are usually about 1.8 times probit
- Predictions are often very close
- Inference are usually the same
- Using the normal distribution is intuitive
- But while log-odds are not intuitive, the link to the simple tabular odds-ratio is attractive

sociology

252