sociology
UNIVERSITY OF LIMERICK

SO5032 Quantitative Research Methods

Brendan Halpin, Sociology, University of Limerick
Spring 2024

Outline

Lecture 0: Course Outline

2024/5 course outline

SO5032 Spring 2024/5 – Module outline

Module Code:	SO5032
Module Title:	Quantitative Research Methods II (MA)
Academic Year:	2024/5
Semester:	Spring
Lecturer(s):	Dr Brendan Halpin
Lecture Locations:	Mon 12-1400 CG055; Lab Tue 12-1400 A0060a
Lecturer(s) Contact Details:	brendan.halpin@ul.ie
Lecturer(s) Office Hours:	Monday 1430-1730

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 χ^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 12-1400	Lab Tue 12-1400
1: Jan 27	Categorical data, association in tables	✓	✓
2: Feb 03	Association in ordinal data	X	✓ (lecture)
3: Feb 10	Understanding multidimensional causality	✓	✓
4: Feb 17	Introducing multiple regression	✓	✓
5: Feb 24	Further multiple regression	✓	✓
6: Mar 03	Multiple regression: residuals & influence	✓	✓
7: Mar 10	Regression with logged dependent variables	✓	✓
8: Mar 17	Introducing logistic regression	X	✓ (lecture)
9: Mar 24	Further logistic regression	✓	✓
10: Mar 31	Multinomial regression	✓	✓
11: Apr 07	Multinomial and ordinal regression	✓	✓
–: Apr 14	Easter break		
12: Apr 21	Ordinal regression continued	✓	✓ (lecture)

sociology

Texts

- Main text: Agresti, *Statistical Methods for the Social Sciences* – particularly chapters 8, 10, 11, 14 and 15
- Supplementary texts:
  - de Vaus, *Surveys in Social Research*: good on survey methodology
  - Agresti, *Introduction to Categorical Data Analysis*
  - Pevalin and Robson, *The Stata Survival Manual*

sociology

Details of Module Assessment:

- Three assignments, weeks 6, 11 and 15.
- The first two assignments are worth 20% each.
- The final assignment is a project, worth 60%, and should be worked on throughout the semester (see below).

sociology

Details of Annual Repeats:

A 100% assignment, to be submitted in the examination period.

sociology

BrightSpace and Other Classroom Technologies:

- The module will use BrightSpace for submission of assignments and for provision of materials.
- `https://teaching.sociology.ul.ie/so5032` may also be used

sociology

IN TERM ASSIGNMENT(S):

- Assignment 1: Homework exercises relating to linear regression.
  - Marks: 20%
  - Deadline: End week 6
- Assignment 2: Homework exercises relating to categorical data analysis.
  - Marks: 20%
  - Deadline: End week 11
- 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.
  - Marks: 60%
  - Deadline: End week 15.

sociology

FEEDBACK:

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.

sociology

Plagiarism notice

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 `https://www.ul.ie/sociology/` under Student Resources).

sociology

Deadline policy

Please also note the Department's policy on deadlines, also available at `https://www.ul.ie/sociology/` under Student Resources.

sociology

Lecture 1: Categorical data analysis

Categorical data analysis

Association between categorical variables

- Association between categorical variables: departure from independence
- Visible in patterns of percentages
- Three main questions (cf Agresti/Finlay p265)
  - Is there evidence of association?
  - What is the form of the association?
  - How strong is the association?

The χ^2 test

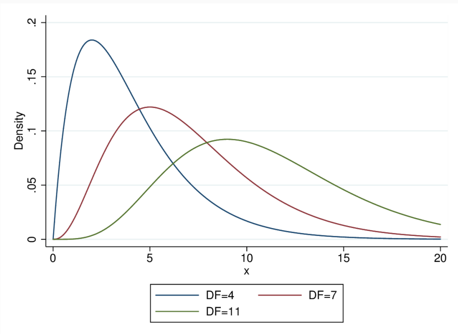
- Compare observed values with expected values under independence:

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

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

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

The χ^2 distribution



Limitations of χ^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 χ^2 statistic

Adjusted Residuals

- The sum of squared Pearson residuals has a χ^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 ($< c5\%$ 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 ∞

Odds ratios

- Odds differ from proportions/percentages:
  - Percentage: $\pi_i = \frac{f_i}{Total}$
  - 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 ∞

Odds ratios

- Odds ratio (i): $\frac{\frac{360}{240}}{\frac{240}{160}} = \frac{1.5}{1.5} = 1$
- Odds ratio (ii): $\frac{\frac{600}{0}}{\frac{0}{400}} = \frac{\infty}{0} = \infty$
- Range: 0 through 1 (no association) to $+\infty$

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"

Ordinal Data

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

Lecture 2

Reading (for this and last week):

- Agresti, Chapter 8

Lecture 2

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

Tabular association in Stata

`tabchi` procedure allows access to

- Percentages
- Expected values
- Residuals
- Adjusted residuals

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

Example: row percentages

Example: observed and expected values

Example: adjusted residuals

Measures of ordinal association

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

Non-linear correlation

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

Truly ordinal measures

- The Gamma statistic (γ) is truly ordinal
- Counts "concordant" and "discordant" pairs

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

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

Gamma in practice

Variants

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

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

Scouting example

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

Scouting example

Low Church Attendance

Scout	Delinquent		
	Yes	No	Total
Yes	10	40	50
No	40	160	200
Total	50	200	250

Medium Church Attendance

Scout	Delinquent		
	Yes	No	Total
Yes	18	132	150
No	18	132	150
Total	36	264	800

High Church Attendance

Scout	Delinquent		
	Yes	No	Total
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:

$$Income = a + b \times PaidWork$$

or

$$Income = a + b \times Female$$

- We can also fit a single multivariate regression

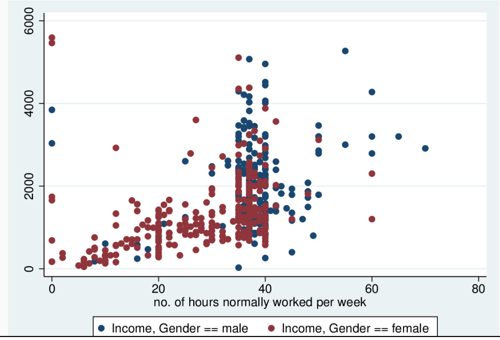
$$Income = a + b \times PaidWork + c \times Female$$

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 $Income = a + b \times PaidWork + c \times Female$ 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

Income, hours and gender



T-test: Income by gender

Regression: Just hours

Regression: Hours and binary gender

Regression: for men only

Regression: for women only

Regression: interaction

Regression: Direct and indirect 1

Regression: Direct and indirect 2

Regression: Direct and indirect 3

Regression: Direct and indirect 4

Outline

- Multiple regression
- Formula, Interpretation
- Hypothesis testing
- Goodness of fit: residuals and R^2
- Agresti, Ch 11

Lecture 4: Summary of multiple regression

Formula

Formula for multiple regression

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

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

- Interpretation of β_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 Xs

Predictions

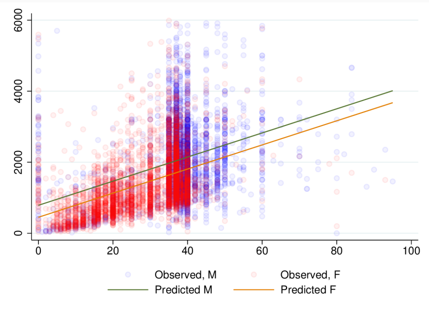
$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_k X_k$

- 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 β_i

Simplest example

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

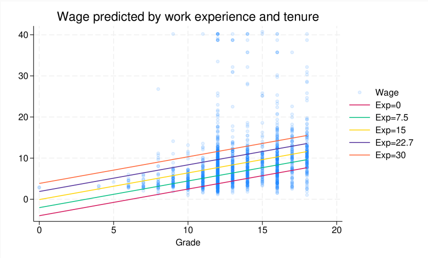
Predicted lines: one for each value of sex



More general 2 X-variable example

Effect of experience on wage, controlling for grade

Effect of grade on wage, controlling for experience



See https://teaching.sociology.ul.ie/so5032/ttlgradelin.html

Residuals

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

$Y = \hat{Y} + e$

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

- Mean of zero
- Standard deviation of σ (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 = $\sum(Y_i - \hat{Y})^2$
- R^2 = $\frac{TSS-RSS}{TSS}$
- Range: 0 (no relationship) to 1 (perfect linear relationship)
- PRE: Proportional Reduction in Error

R^2 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: $abs(\hat{\beta}_j/\mathrm{se}_j) > t$
- Interpretation:
  - Null: population value of β is 0; this variable has no influence once the other variables are taken account of

Example

Hypothesis testing: all parameters together

- F-test:
  - $\beta_1 = \beta_2 \ldots = \beta_k = 0$
- 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: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_g X_g$
  - Model 1: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_g X_g + \beta_h X_h ... + \beta_k X_k$
- Null hypothesis: $\beta_h = \ldots = \beta_k = 0$
- 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.

More than two categories

With more that 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

Interactions

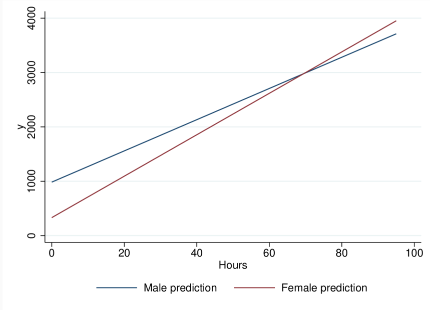
- An interaction effect is where the effect of one variable on Y changes depending on the value of another

Income, hours and gender

For men

For women

Different effects



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 $\beta_1 + \beta_3 X2$ increase in $\hat{Y}$
- Equivalently, a 1-unit increase in X_2 leads to a $\beta_1 + \beta_3 X1$ increase in $\hat{Y}$

Interaction between hours and sex

- Simplest example: one variable is binary

$$\hat{Y}_m = \beta_0 + \beta_1 X_1 + \beta_2 \times 0 + \beta_3 X_1 \times 0$$
$$\hat{Y}_f = \beta_0 + \beta_1 X_1 + \beta_2 \times 1 + \beta_3 X_1 \times 1$$

One-unit increase

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

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

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
```


Results

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

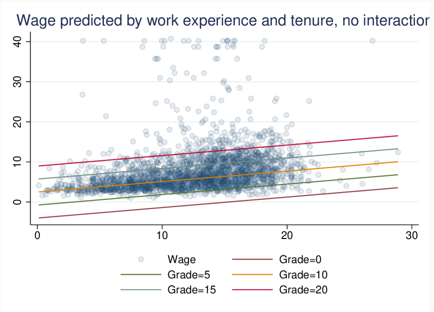
Same results using Stata's formula syntax

Predictions

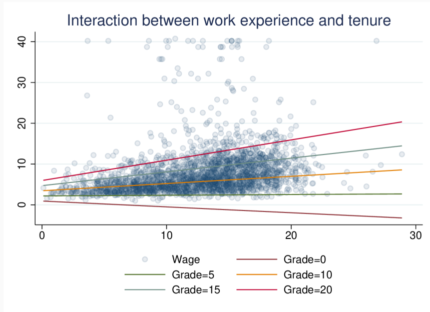
Sex	Hrs	β_0	β_1	β_2	β_3	$\hat{y}$
M	0	983.9722	+ 0*28.71923	+ 0*-653.2448	+ 0*0*9.399515	= 983.9722
M	80	983.9722	+ 80*28.71923	+ 0*-653.2448	+ 80*0*9.399515	= 3281.5106
F	0	983.9722	+ 0*28.71923	+ 1*-653.2448	+ 0*1*9.399515	= 330.7274
F	80	983.9722	+ 80*28.71923	+ 1*-653.2448	+ 80*1*9.399515	= 3380.227

Interactions between two continuous variable

Without interaction, predictions for different levels of grade



With interaction



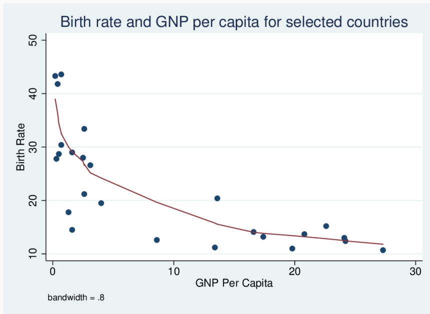
Lecture 5: Interaction and Non-linearity

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")
```


Nonlinear plot

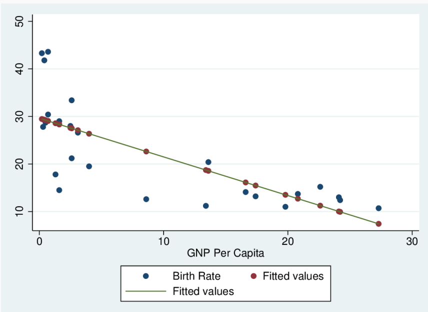


Get linear relationship

```
reg bir gnp

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


Linear plot



Quadratic

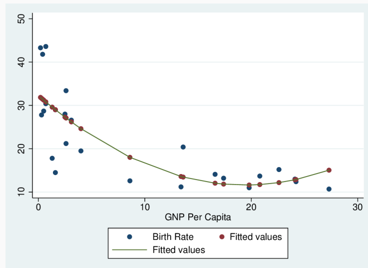
Linear regression doesn't fit well

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

Let's try quadratic:

Quatratic plot

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

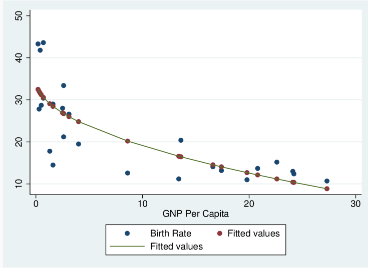


$\sqrt{GNP}$

Let's try square root of GNP:

$\sqrt{GNP}$ plot

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

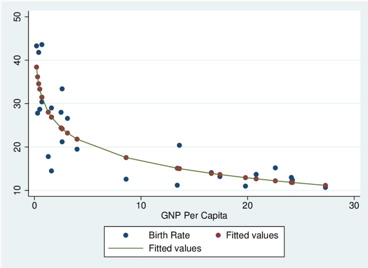


log(GNP)

Let's try the log of GNP:

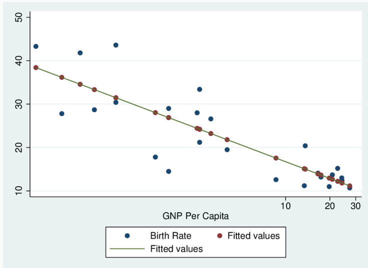
log(GNP) plot

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



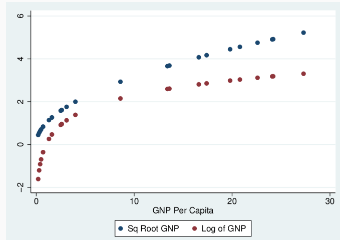
Log-scale plot

```
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
```



Residuals

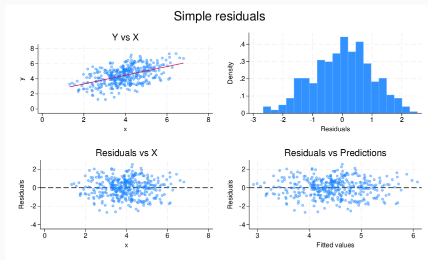
$Y = b_0 + b_1 X_1 + ... + b_k X_k + e$

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

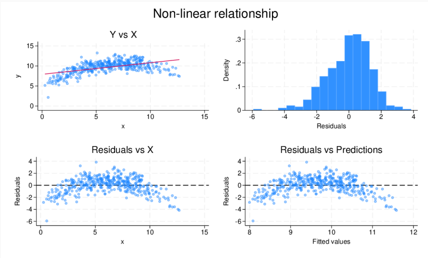
Characteristics

- 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

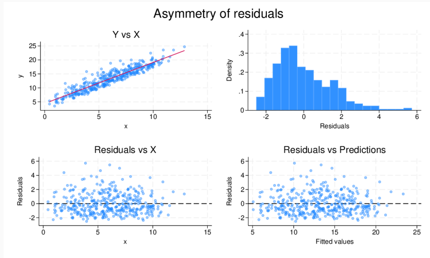
Examining residuals: ideal



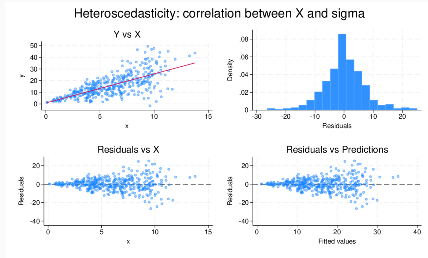
Examining residuals: Non-linear



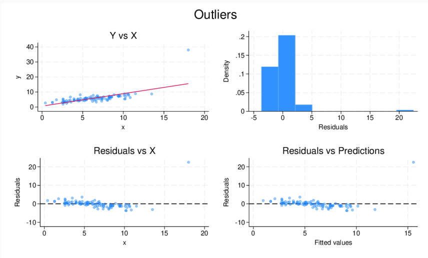
Examining residuals: asymmetric



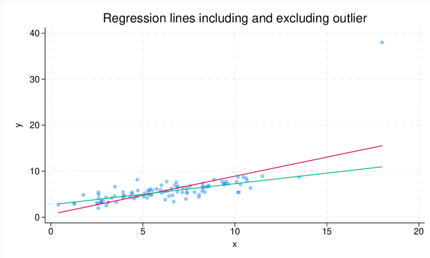
Examining residuals: heteroscedasticity



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

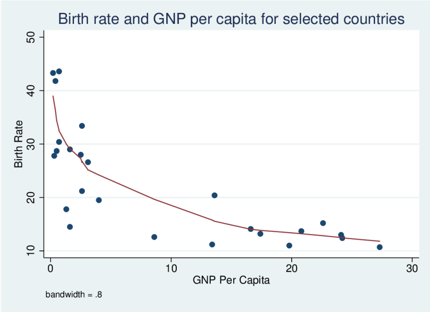
Outlier interactive app

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

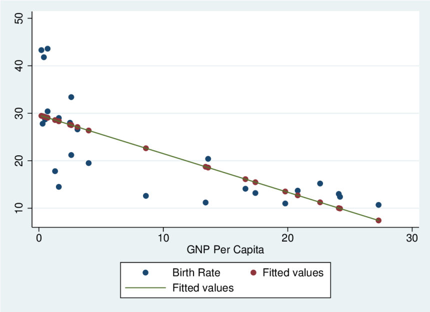


Get linear relationship

```
reg bir gnp

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


Linear plot



Quadratic

Linear regression doesn't fit well

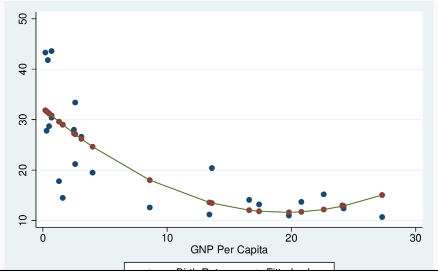
Clearly, as GNP rises BIR falls, but the rate of fall declines

Let's try quadratic:

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


Quatratic plot

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



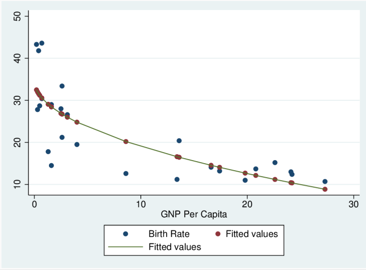
$\sqrt{GNP}$

Let's try square root of GNP:

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


$\sqrt{GNP}$ plot

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



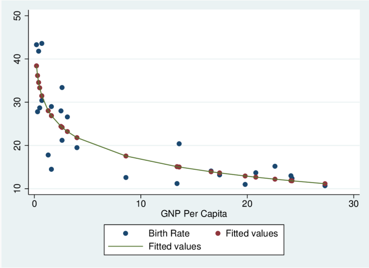
log(GNP)

Let's try the log of GNP:

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

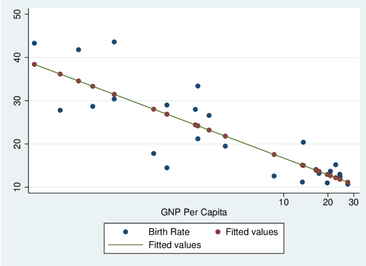

log(GNP) plot

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



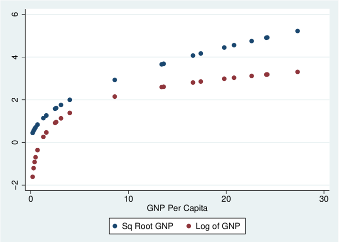
Log-scale plot

```
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:

log(10) = 1	$10^1 = 10$
log(100) = 2	$10^2 = 100$
log(1000) = 3	$10^3 = 1000$
log(1000000) = 6	$10^6 = 1000000$

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

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

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:

100 × 1000 = 100000

$10^2 \times 10^3 = 10^5 = 10^{2+3}$

Calculate A × B

Thus do calculate A × B we do as follows:

- Calculate log(A)
- Calclate 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 α (i.e., a 10% increase means α = 1.1)

Compound interest

Year 0	100
Year 1	100 × α
Year 2	100 × α × α
Year 3	100 × α × α × α
Year 4	100 × α × α × α × α
Year 5	100 × α × α × α × α × α

In short, the value in year t is 100×α^t

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

Constant proportional increase

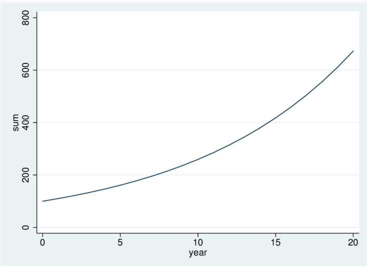


Figure 1: A constant proportional increase

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 α every year, we add log(α).

Plot

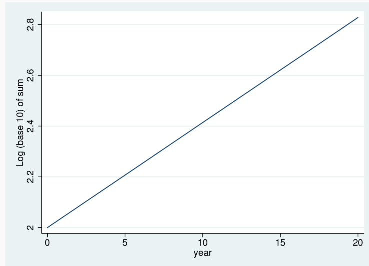


Figure 2: Taking the base-10 log of the sum: a straight line

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

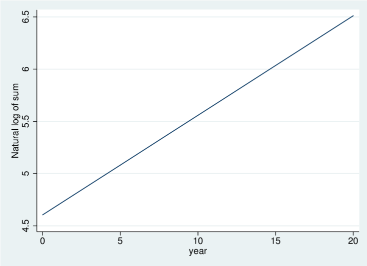


Figure 3: Taking the natural log of the sum: also a straight line

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)

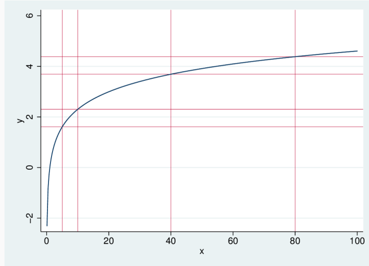


Figure 4: The natural log of X for X from 0.1 to 100

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.

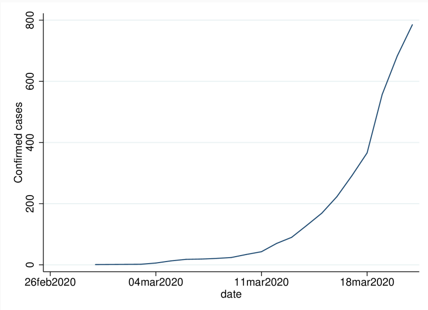
Confirmed cases in Ireland

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



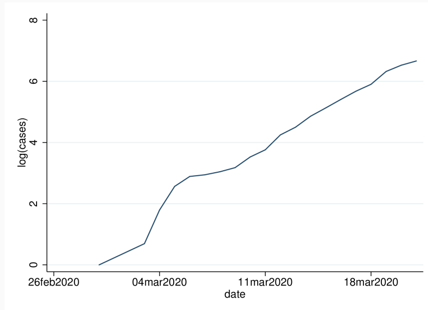
Log cases

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 very 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!

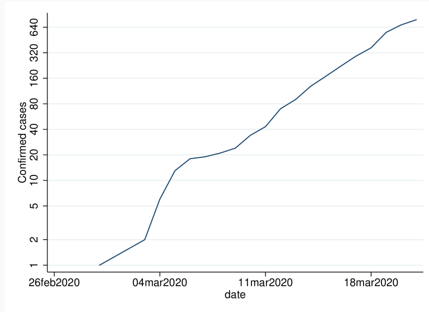
Log scale, real cases

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

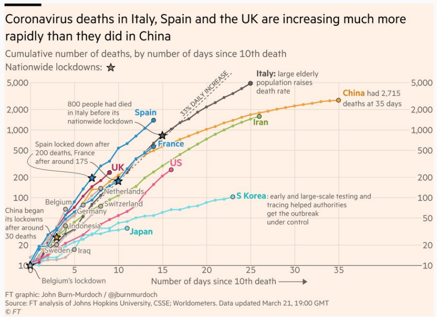
```
line cases date, yscale(log) ylabel(1 2 5 10 20 40 80 160 320 640)
```

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



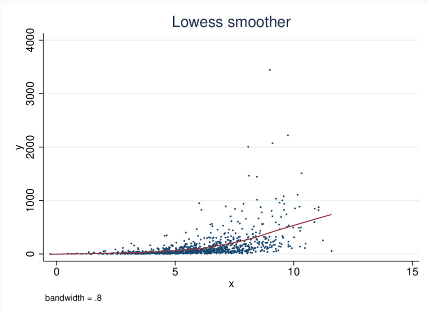
Lecture 7: Logs and log regression

Log regression

Multiplicative relationship

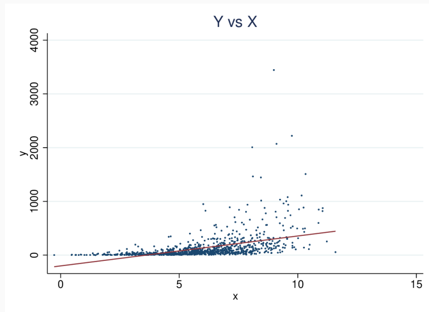
- Where the underlying relationship is multiplicative, linear regression doesn't work well
- Implies an additive increase where a multiplicative one is better
- If we take the log of the dependent variable:
  - better estimates
  - often cures heteroscedasticity

Simulation: Y increases 65% for X +1



Linear regression

Predictions

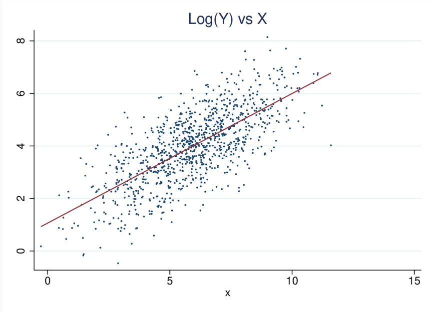


Log(Y)

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

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

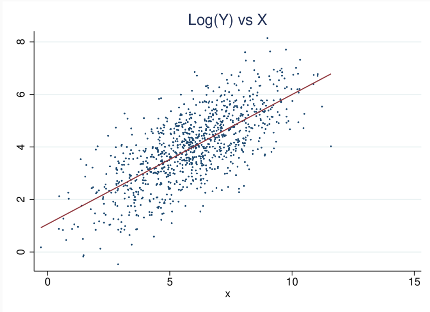
- where RMSE is the standard deviation of the regression

Calculations

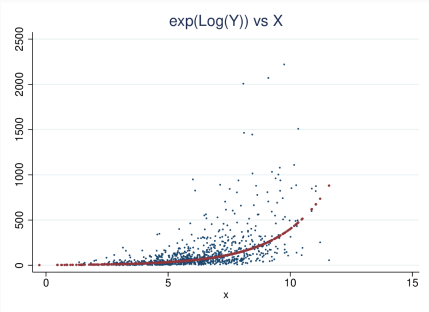
```
gen ly = log(y)
reg ly x

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

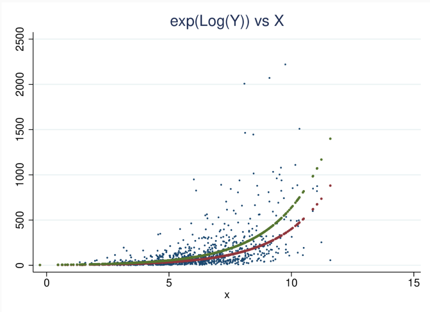

Predictions: predict log(Y) on log scale



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



Predictions: with correction



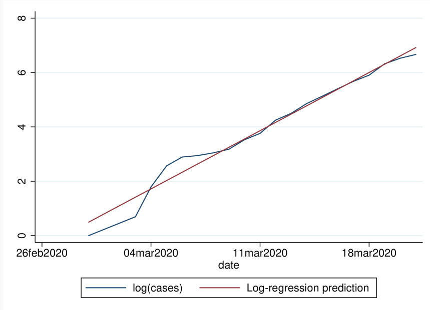
Predicting COVID-19

- 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(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

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

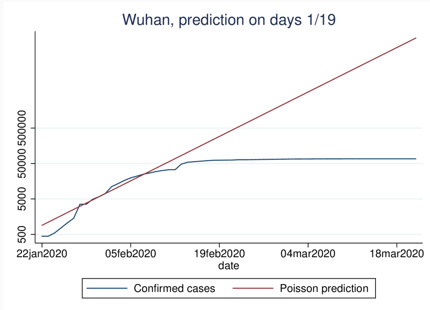
Exponential increase cannot go on indefinitely

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

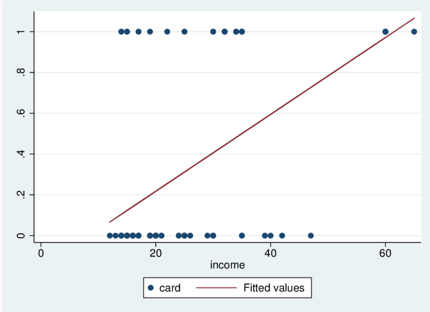
- If we use OLS with binary outcomes, it is called “linear probability model”:

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

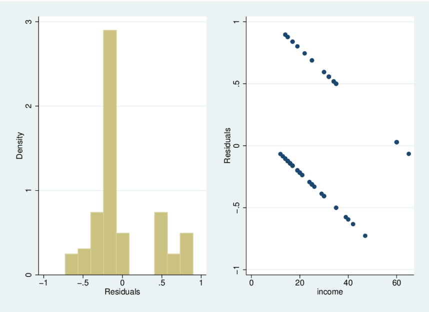
- data is 0/1, prediction is probability
- Assumptions violated, but if predicted probabilities in range 0.2–0.8, not too bad

Credit card example

Credit card example



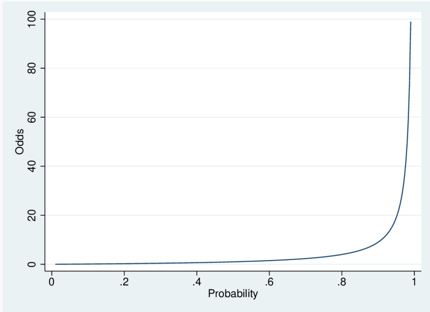
Credit card example



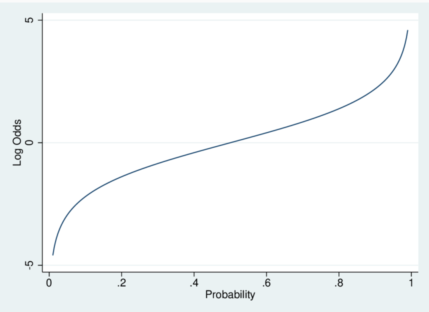
Logistic transformation

- 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 $-\infty : \infty$

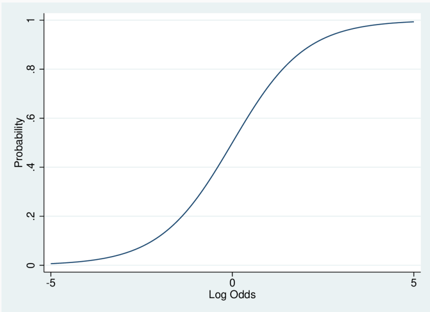
Probability to odds



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:

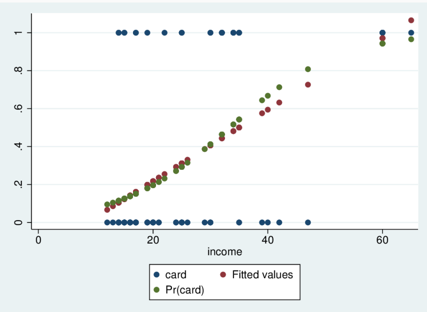
$$Pr(Y=1) = \frac{e^{a+bX}}{1+e^{a+bX}} = \frac{1}{1+e^{-a-bX}}$$

Parameters

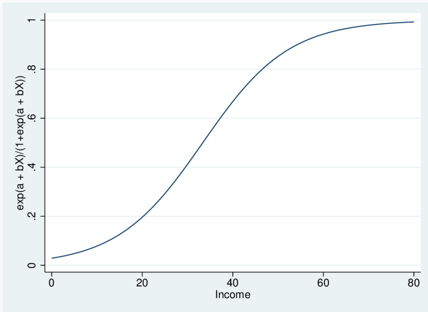
- The b parameter is the effect of a unit change in X on $\log\left(\frac{Pr(Y=1)}{1-Pr(Y=1)}\right)$
- This implies a multiplicative change of e^b in $\frac{Pr(Y=1)}{1-Pr(Y=1)}$, in the Odds
- Thus an odds ratio
- But the effect of b on P depends on the level of b

Credit card logistic regression

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 odds/(1+odds) gives us the probability

Calculating predicted probabilities

- Example: log(odds) = 0.25 + 0.12X
- Predict for X == 10
  - Predicted log-odds = 0.25 + 0.12*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/