REGRESSION MODELLING

DATA SOURCES

- Where did the data come from?
- Ask the following of the data:
  - Was the explanatory variable manipulated?
- Two answers:
- No: therefore it is observational: data gathered is observed with no interference by researcher.
- Yes: therefore it is experimental: data gathered is manipulated by the researcher.

DATA SOURCES

- Observational data typically generated through data reporting, this is collecting and organising data to monitor a process or phenomenon. No particular hypothesis in mind. Although it is often analysed later to test specific hypothesis. Data reporting tells you what happened, data analysis tells you why.
  - Examples are Gapminder, Nesarc, AddHealth, theses are either reporting or survey
- Experimental data comes from studies in which groups of observations are either pre-selected or randomly assigned and the values of an explanatory variable are observed on some response variable. Usually one explanatory variable is manipulated and all other variables are held constant. There is also usually a control group. Observations must be randomly assigned to values of the explanatory variable. E.g. growth of plants. Conclusions can be drawn.

CONFOUNDING VARIABLES

- The greatest weakness of observational studies are confounding variables.
- Suppose a study is conducted on quitting smoking. We are trying to determine which method is most effective, drugs for nicotine addiction, therapy, the combination of both, or simply quitting.
- The explanatory variable is the method. The response variable is the success or failure of quitting.
- The study shows that the combination of both therapy and drugs was most successful. There is clear evidence of an association between this method and success of quitting.

CONFOUNDING VARIABLE

- Can we then conclude that the combination of drug and therapy method causes success more than using either drugs or therapy on their own?
- In this study the subjects opted for the method. Could it be that some other variable is impacting the response? Such as older people are more likely to choose certain methods of quitting? Perhaps older people are more successful in quitting than younger people?
- The data makes it appear that the method alone was responsible for the success of quitting however it may be some other reason.
- This other reason is a confounding variable.

CONFOUNDING VARIABLES

- We could control the confounding variable for example in this case Age by studying older and younger adults separately. Then if both older and younger adults who choose one method have a higher success rate then we would be closer to producing evidence of causation.
- Why is causation so important?
- Controlling age would not allow us to make a definite claim of causation. Why?
- What about motivation to quit? Desire to quit? Income? Gender?

CONFOUNDING VARIABLES

- Observational studies cannot prove causality.
- They are however an extremely common tool used to draw conclusions about causal connection. To do this we must try to control confounding variables as much as possible.
- Association does not imply causation.
- Association between wildfires and fire fighters deployed….
- Confounding is a major threat to the validity of inferences made about statistical associations. We test for confounders by including these third variables, or fourth, fifth and sixth, in our statistical models that may explain the association of interest.
- We want to demonstrate that our association of interest is significant even after controlling for potential cofounders.

MULTIVARIATE METHODS

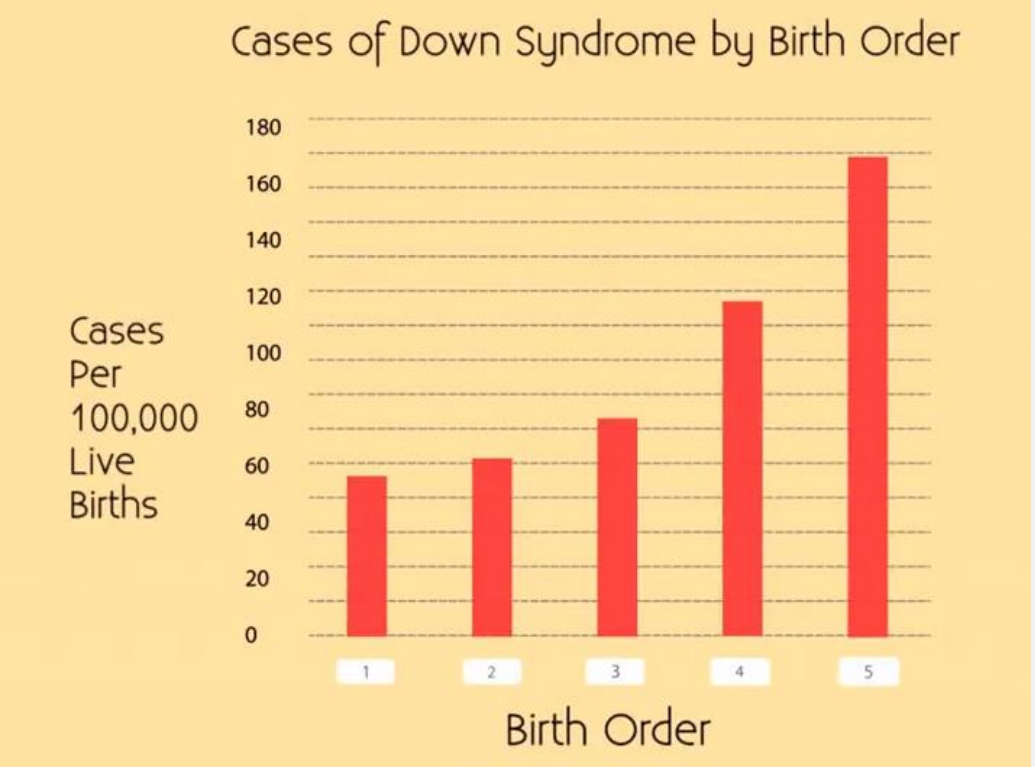
- Adding potential confounding variables to a statistical model can help us to gain a deeper understanding of the relationship between variables or lead us to rethink an association.
- Multivariate methods are commonly used to learn more about the relationship between multiple explanatory variables and a response variable.
- E.G.
- What are the best predictors of success in college?
- What factors predict caterpillar reproduction?

MULTIVARIATE METHODS

- Lurking variables may be tied in with the explanatory variable.
- You may have already identified a significant relationship between your explanatory and response variables.
- Now you want to check if the relationship is real.
- For example: the association between birth order and the number of cases of down syndrome per 100,000 live births.

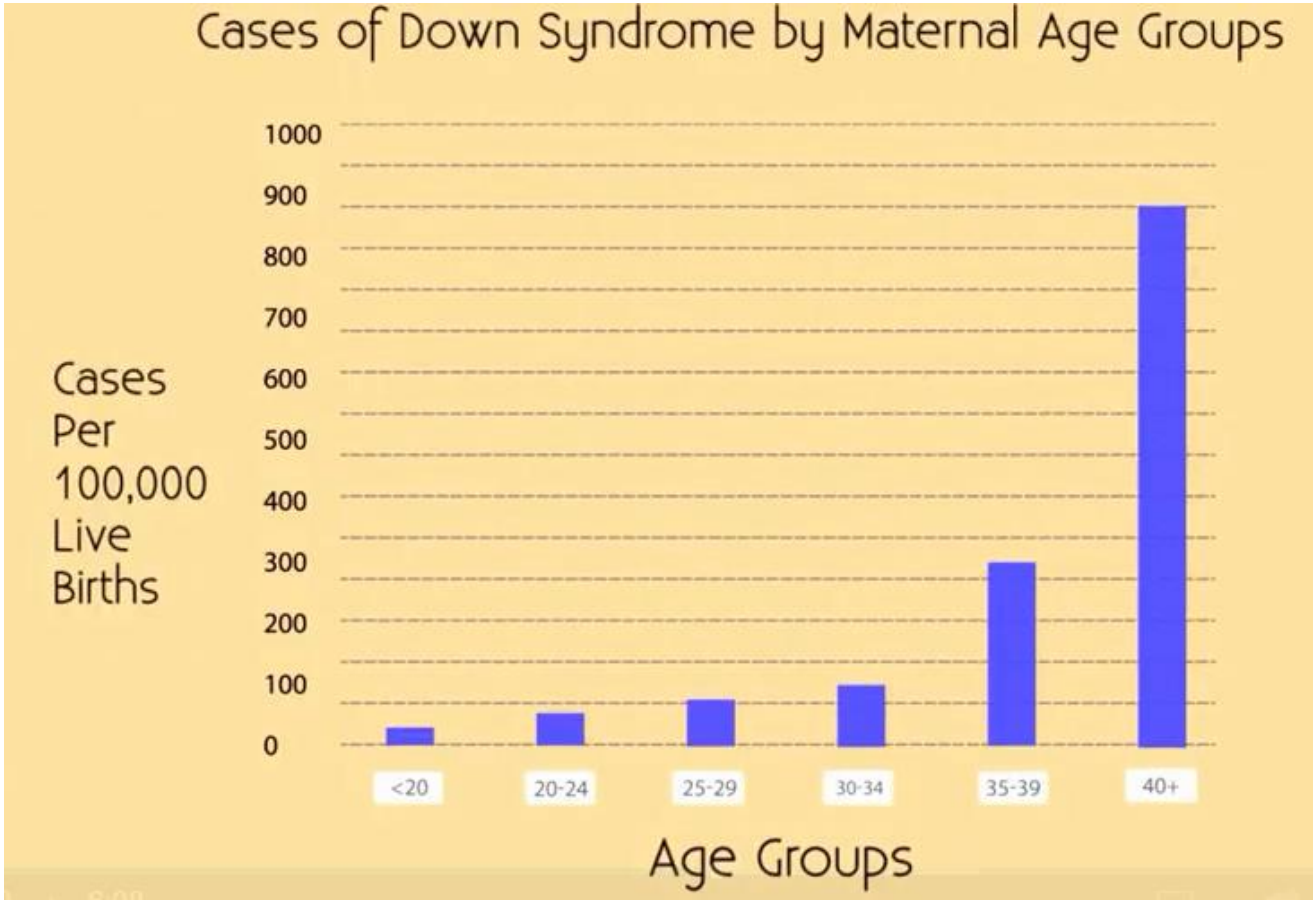
REGRESSION MODELLING

- It looks like a linear association where the first born in a family has the lowest likelihood of having down syndrome.
- With fifth born children having an increased risk.
- This relationship would also show as statistically significant using chi square test of independence.
- However…



REGRESSION MODELLING

- Another statistically significant relationship is the association between maternal age at a child's birth and the likelihood that the child will have down syndrome.
- Here there are really low rates for mothers up to 29, and 35-39 and older rates are clearly higher.

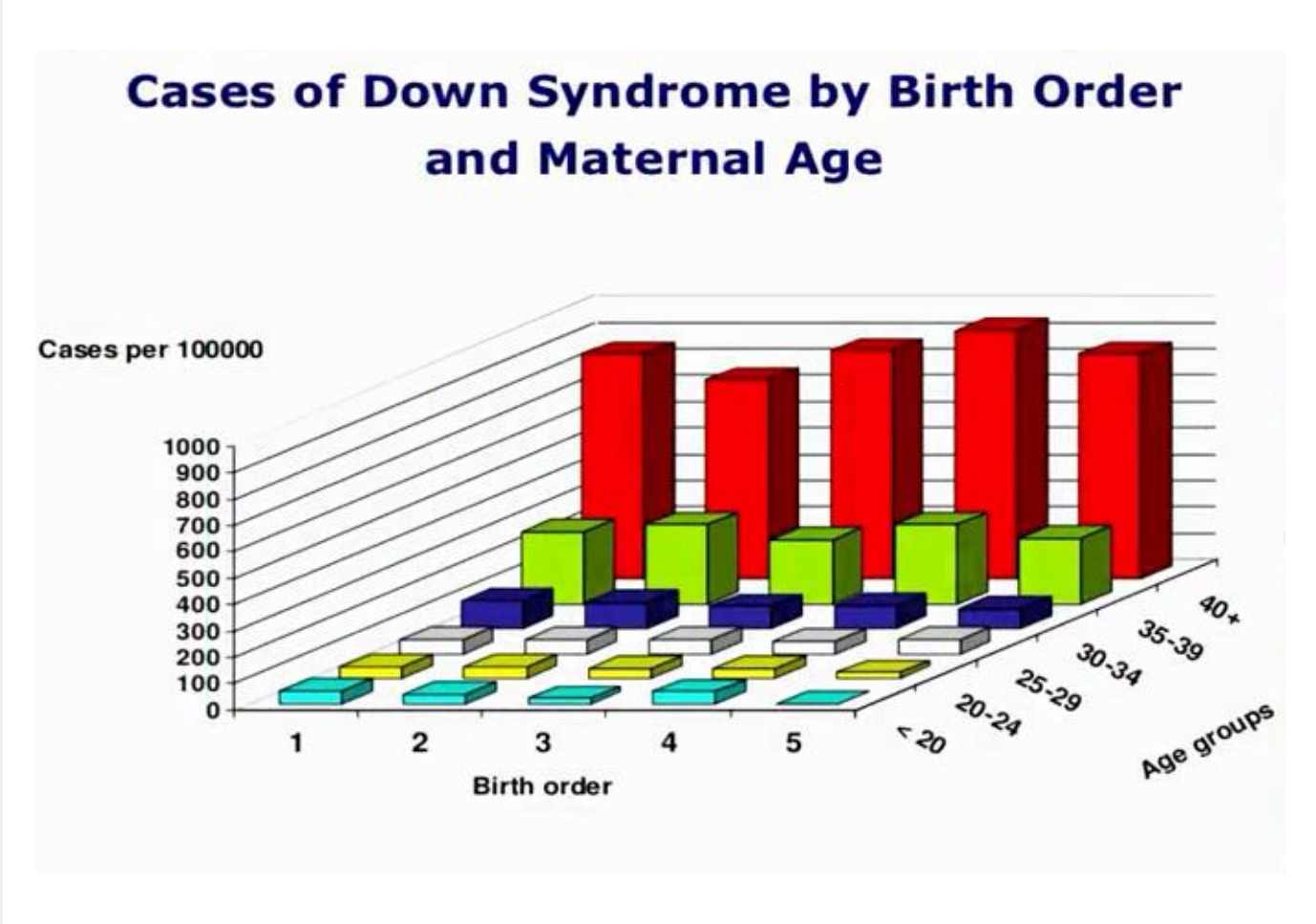


REGRESSION MODELLING

- We test for cofounders by including these third or fourth variables in our statistical models that might explain the association interest.
- In this example it is possible that the association between a child's birth order and the risk of down syndrome could be confounded by maternal age.
- Alternatively, the association between maternal age and down syndrome might be confounded by birth order.
- It may be that after controlling for each variable that both independently predict the likelihood of a diagnosis.
- We can chart to show which is the case.

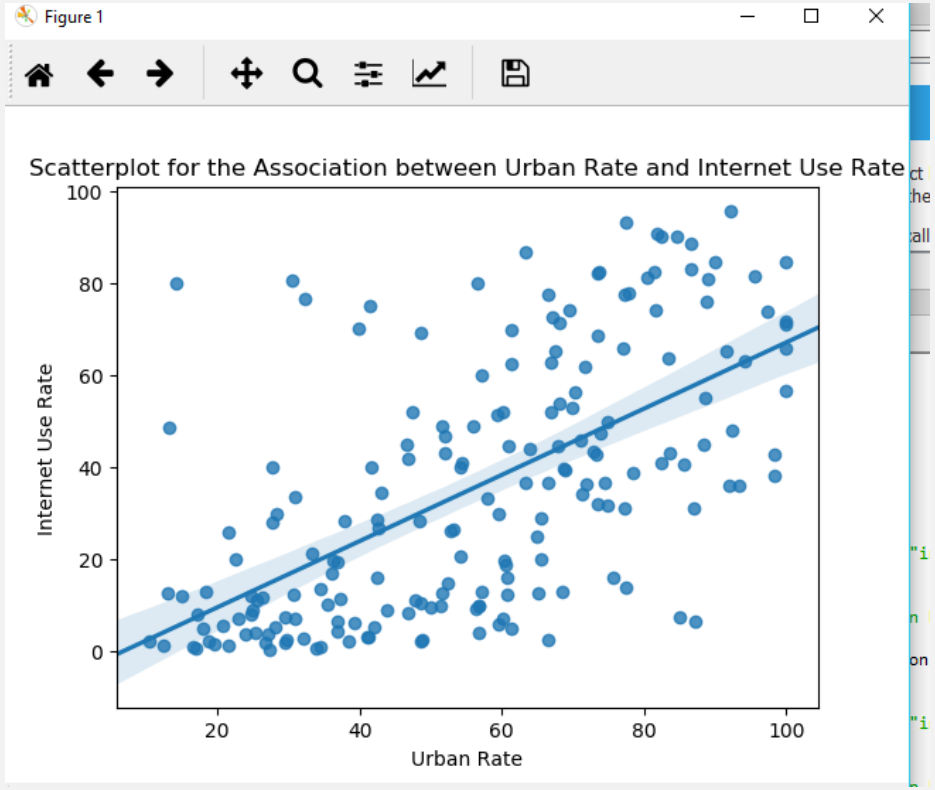
REGRESSION MODELLING

- This graph tells us that maternal age is associated with Down Syndrome.
- Birth order does not show a significant relationship with diagnosis as there is really no difference.
- Maternal age however shows an upward trend as maternal age rises.
- Once we control for birth order the relationship still exists between diagnosis and maternal age. Birth order does not confound the relationship between maternal age and down syndrome.



REGRESSION MODELLING

- We must use multiple regression to evaluate multiple explanatory variables., and or potential confounders when predicting a quantitative response variable.
- How does linear regression analysis work?
- The graph we created in step 2 of Lab 8 showed the relationship between urbane population rate and the internet use rate.
- Explanatory variable urban rate on X axis, response variable internet use rate on Y axis.
- If we ran a correlation coefficient it would give us .61, a pretty strong positive linear association.



REGRESSION MODELLING

- Next we test the model. To do this we must determine the equation of that best fit line.
- The line in the graph shows the best fit and is defined as:
  - $y = mx + b$
  - x is the variable and y are the variables on the axis.
  - m is the slope
  - b is the spot on the Y-intercept where the line crosses an axis.
- In Python we use the ordinary less squares or OLS function from the stats model formula API library

REGRESSION MODELLING

```
modelname = smf.ols('quant_response ~
quant_explanatory', data=dataframe).fit()


Print(modelname.summary())
```

REGRESSION MODELLING

- Linear regression is one of the simplest and most commonly used modelling techniques.
- It makes very strong assumptions about the relationship between the explanatory variables (X) and the response variable (Y). It assumes that this relationship takes the form:
- $\hat{Y} = ß_0 + ß_1 * x$
- Ordinary least squares is the simplest and most common estimator in which the two "beta's" (ß) are chosen to minimise the square of the distance between the predicted values and the actual values.
- This method is easier to interpret than more sophisticated models and can provide insight into what the model captures.

REGRESSION MODELLING

```
g5 = smf.ols('NDSymptoms ~ DYSLIFE + MAJORDEPLIFE + numcigsmoked_c + age_c + SEX", data=sub1).fit()
print (reg4.summary())
```

OLS Regression Results

Dep. Variable:	NDSymptoms	R-squared:	0.136
Model:	OLS	Adj. R-squared:	0.133
Method:	Least Squares	F-statistic:	41.08
Date:	Fri, 23 Oct 2015	Prob (F-statistic):	2.39e-39
Time:	13:56:07	Log-Likelihood:	-2591.2
No. Observations:	1313	AIC:	5194.
Df Residuals:	1307	BIC:	5226.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	2.2550	0.156	14.480	0.000	1.949	2.561
DYSLIFE	0.2746	0.209	1.316	0.188	-0.135	0.684
MAJORDEPLIFE	1.2975	0.116	11.161	0.000	1.069	1.526
numbercigsmoked_c	0.0353	0.006	6.257	0.000	0.024	0.046
age_c	-0.0400	0.022	-1.806	0.071	-0.083	0.003
SEX	-0.0439	0.099	-0.442	0.658	-0.238	0.151

Omnibus:	69.558	Durbin-Watson:	2.075
Prob(Omnibus):	0.000	Jarque-Bera (JB):	48.596
Skew:	0.361	Prob(JB):	2.80e-11
Kurtosis:	2.394	Cond. No.	38.5

REGRESSION MODELLING

- Remember that our data is from a sample of the population, therefore our variables in our regression model are called parameter estimates. They are only estimates of what the population values may be.
- If we were to take a new sample and perform the same regression model on the new sample, the parameter estimates are not likely to be the same as they were with our first sample. This is due to sampling variability.
- The two last columns in the output show us the lower and upper limits of the parameter estimates at the 95% confidence interval.

CONFIDENCE INTERVALS

- 95% certain that the true population parameter for the association between major life depression and number of nicotine dependant symptoms fall somewhere between 1.1 and 1.5.
- That is, in the population, there's a 95% chance that people with major life depression have anywhere between 1.1. and 1.5 more nicotine dependent symptoms than people without major depression.