Economic Evaluation alongside natural experiments

Heather Brown

Using observational data for economic evaluations

- Identify appropriate linked datasets

Using observational data for economic evaluations

- Put the data into context
- How does the study fit into the local/political context-justify your choice of data
- Why is your data the best for your question
- How much pre-treatment data you have

Using observational data for economic evaluations

- How do you choose an appropriate comparison group?
- Consider multiple comparison groups

Using observational data for economic evaluations

- Methodologies to control for selection bias

1. Sampling Bias
2. Survivorship Bias
3. Exclusion Bias
4. Volunteer or Self-selection Bias
5. Attrition Bias
6. Recall Bias

Using observational data for economic evaluations

- Measurement Error (due to differences in timing because of the intervention and data)
- How could this bias your data? How can you reduce this bias?

Using observational data for economic evaluations

- Incorporating externalities
- Spatial spillovers
- Are there any other relevant interventions going on at the same time?

Using observational data for economic evaluations

- Are you planning on exploring equity issues
- What sub-groups will you look at?
- Potential behavioural responses to interventions (have they been identified and can they be measured)

Using observational data for economic evaluations

- Decide on an economic evaluation technique
- If using Cost utility analysis-can you map an intermediate outcomes to QALYs? (potentially using other data for utility values)
- Difficulty with unidimensional measures

Using observational data for economic evaluations

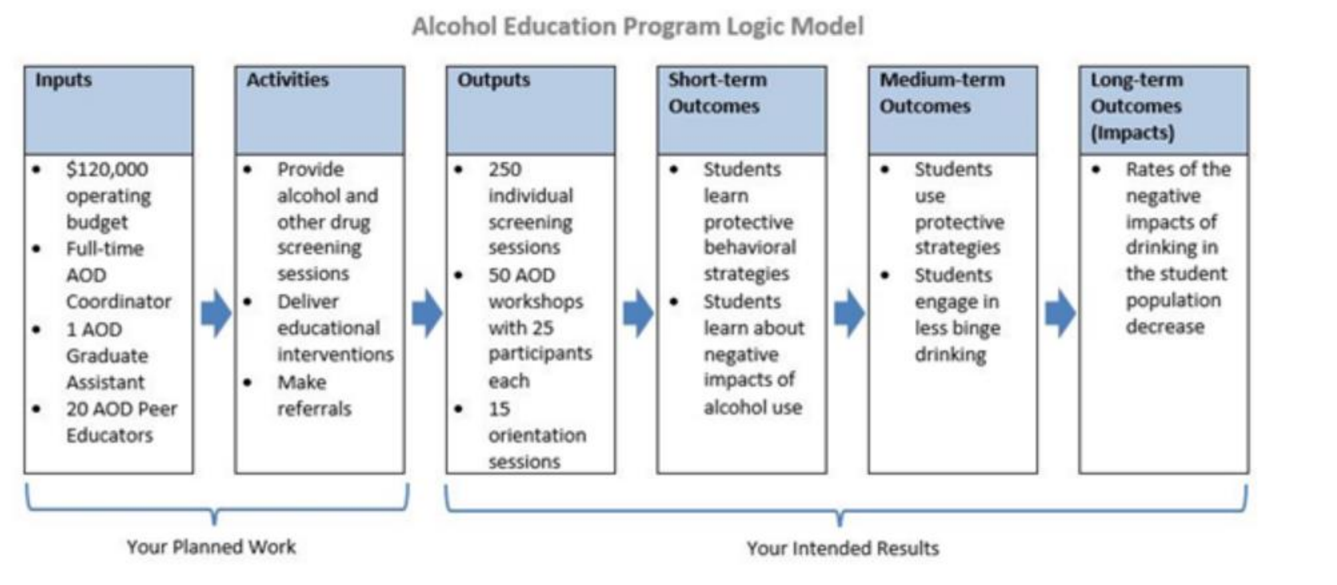
- Costs (which perspective to take)
- Unit costs vs average unit cost of most frequently used service-justification of which costs are used and for what reason
- What to do if you don't have costs for a specific element

Using observational data for economic evaluations

- Time Horizon
- Discount Rate

Using observational data for economic evaluations

- Develop a logic model



Alcohol Education Program Logic Model

Inputs	Activities	Outputs	Short-term Outcomes	Medium-term Outcomes	Long-term Outcomes (Impacts)
• $120,000 operating budget • Full-time AOD Coordinator • 1 AOD Graduate Assistant • 20 AOD Peer Educators	• Provide alcohol and other drug screening sessions • Deliver educational interventions • Make referrals	• 250 individual screening sessions • 50 AOD workshops with 25 participants each • 15 orientation sessions	• Students learn protective behavioral strategies • Students learn about negative impacts of alcohol use	• Students use protective strategies • Students engage in less binge drinking	• Rates of the negative impacts of drinking in the student population decrease

Your Planned Work

Your Intended Results

From: http://studentaffairsassessment.org/entries/blog/logic-models

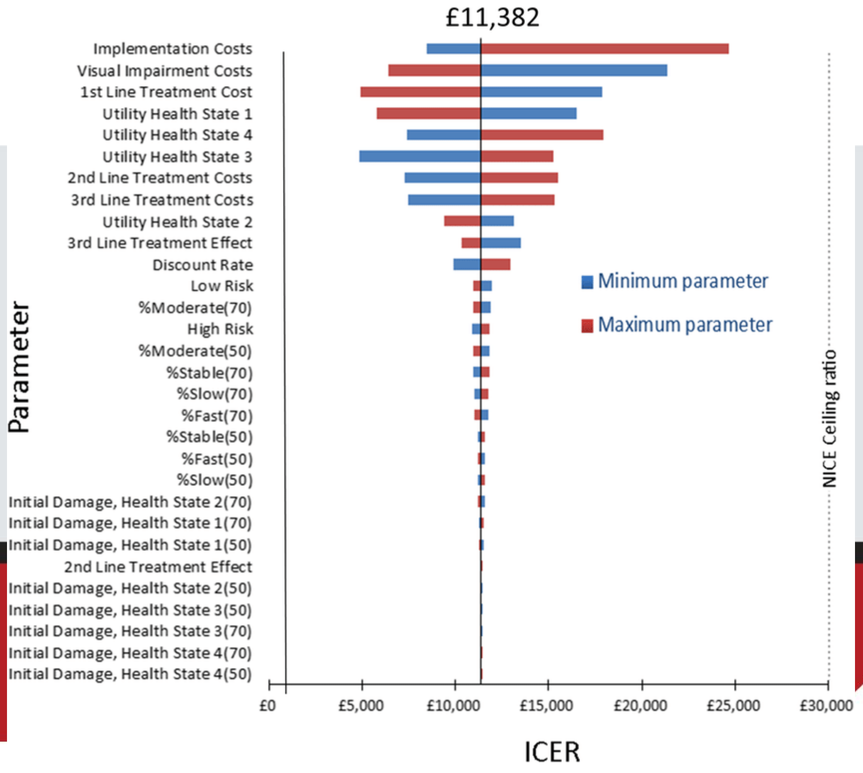
Using observational data for economic evaluations

- Choosing the right estimation model
- How can you account for variation in exposure to intervention in treatment group
- The methodology to reduce bias fits within economic evaluation frameworks
- Controls/Confounders (do you include-if so how do you decide what to control for)

Using observational data for economic evaluations

- Uncertainty and sensitivity analysis
  - – Probabilistic sensitivity analysis
  - – Tornado Diagrams

From: Boodhna, T., & Crabb, D. P. (2016). More frequent, more costly? Health economic modelling aspects of monitoring glaucoma patients in England. *BMC Health Services Research*, *16*(1), 1-13.



Using observational data for economic evaluations

- Report results from all model specification
- Sensitivity analysis

Reference

- Deidda, M., Geue, C., Kreif, N., Dundas, R., & McIntosh, E. (2019). A framework for conducting economic evaluations alongside natural experiments. *Social science & medicine*, *220*, 353-361.

- How to use econometrics for intervention development

Research question (Step 1)

- Suppose you are asked what would be the best policy to prevent rising obesity rates.
- You have been tasked with investigating if any interventions can be developed in relation to schooling.
- You think that those with more education will be less likely to be obese.
- This is based on the Grossman model (Grossman 1972)

Estimation

- What model structure you use depends on your data and research question
- Some options are:

1. Ordinary Least Squares
    - Simplest – basic forms can be done with pen and paper
2. Generalised Least Squares (Random Effects)
3. Fixed effects
4. Binary Probability Models (probit/logit)

Data (Step 2)

- We are going to use data from waves 6-9 (2006-2009) of the Household Income and Labour Dynamics of Australia (HILDA) survey.
- It is a nationally representative survey of households in Australia which began in 2001.
- All household members over the age of 15 are interviewed on an annual basis.
- More information about the data can be found on: http://www.melbourneinstitute.com/hilda/

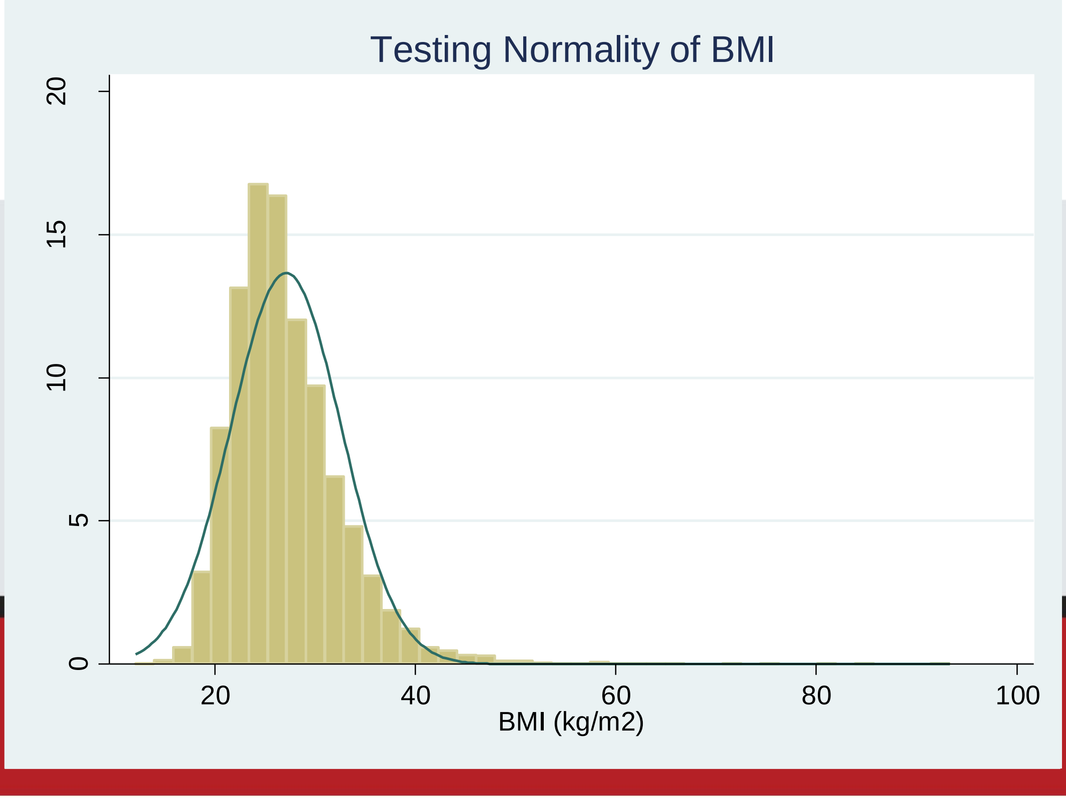
Descriptive Statistics (Step 3)

Lancaster University

Variable	Obs	Mean	Std. Dev.	Min	Max
BMI2	22270	27.11	5.52	12.1	93.3
age	24987	44.65	11.06	25	65
female	24987	0.52	0.50	0	1
highschool	24877	0.12	0.33	0	1
cert1_2	24877	0.01	0.12	0	1
cert3_4	24877	0.23	0.42	0	1
diploma	24877	0.10	0.30	0	1
degree	24877	0.28	0.45	0	1
postgrad	24877	0.12	0.33	0	1
disadvanta~d	24984	0.27	0.44	0	1
loghhincome	24860	10.30	0.71	4.65	13.74
smokes	22914	0.21	0.41	0	1
frequent_pa	22985	0.50	0.50	0	1
married	24979	0.62	0.49	0	1
employed	24987	0.77	0.42	0	1
unemployed	24987	0.02	0.15	0	1

Normal Distribution (Step 4)

- Is the dependent variable normally distributed?



Testing for multicollinearity (Step 5)

Shows correlation between age and female

	age	female	highsc~l	cert1_2	cert3_4	diploma	degree	postgrad
age	1							
female	0.00	1						
highschool	-0.10	0.04	1					
cert1_2	0.02	0.02	-0.04	1				
cert3_4	-0.01	-0.18	-0.20	-0.06	1			
diploma	0.01	0.01	-0.13	-0.04	-0.18	1		
degree	-0.12	0.04	-0.23	-0.08	-0.34	-0.21	1	
postgrad	0.00	0.02	-0.14	-0.04	-0.20	-0.13	0.59	1
disadvanta~d	0.00	0.00	0.01	0.04	0.02	-0.05	-0.14	-0.11
loghhincome	-0.02	-0.06	-0.03	-0.05	-0.07	0.03	0.28	0.20
smokes	-0.11	-0.08	0.02	0.04	0.06	-0.03	-0.16	-0.11
frequent_pa	0.03	-0.05	-0.01	0.00	0.01	0.01	0.01	0.04
married	0.11	-0.02	-0.01	-0.01	-0.02	0.02	0.04	0.04
employed	-0.23	-0.18	0.00	-0.03	0.05	0.02	0.14	0.09
unemployed	-0.04	0.00	-0.01	0.03	0.01	0.01	-0.03	-0.03

Shows correlation between smoking status and log of household income

	disadv~d	loghhi~e	smokes	freque~a	married	employed	unempl~d	
disadvanta~d	1							
loghhincome	-0.19	1						
smokes	0.13	-0.10	1					
frequent_pa	-0.03	0.08	-0.04	1				
married	-0.11	-0.04	-0.21	-0.03	1			
employed	-0.12	0.33	-0.02	0.02	0.03	1		
unemployed	0.05	-0.10	0.06	0.00	-0.06	-0.27	1	

Choose a model specification (Step 6)

You start by deciding to estimate the following model:

$$BMI_{it} = \alpha + \beta_1 Individual_{it} + \beta_2 Household_{it} + \beta_3 Health_{it} + \beta_4 Education + \varepsilon_{it}$$

- You estimate this model using Ordinary Least Squares

Ordinary Least Squares

- Zero mean value of ε: $E(\varepsilon | X_{1,} X_2, X_3)=0$

Mean of the error term is equal to zero. Thus, it shouldn't affect your results.

- No serial correlation between error terms

$$cov(\varepsilon_i, \varepsilon_j)=0, \quad i \neq j$$

Error term from data collected this year is independent of the error term on data collected last year

- Homoscedasticity:

$$\mathrm{var}(Y_i) = \sigma^2$$

The spread/variance of the dependent variable is the same for all explanatory variables.

Ordinary Least Squares

- Zero covariance between ε_i and each *X* variable

$$\operatorname{cov}(\varepsilon_i, X_{2i}) = \operatorname{cov}(\varepsilon_i, X_{3i}) = 0$$

There is no correlation between the error term and the explanatory variables

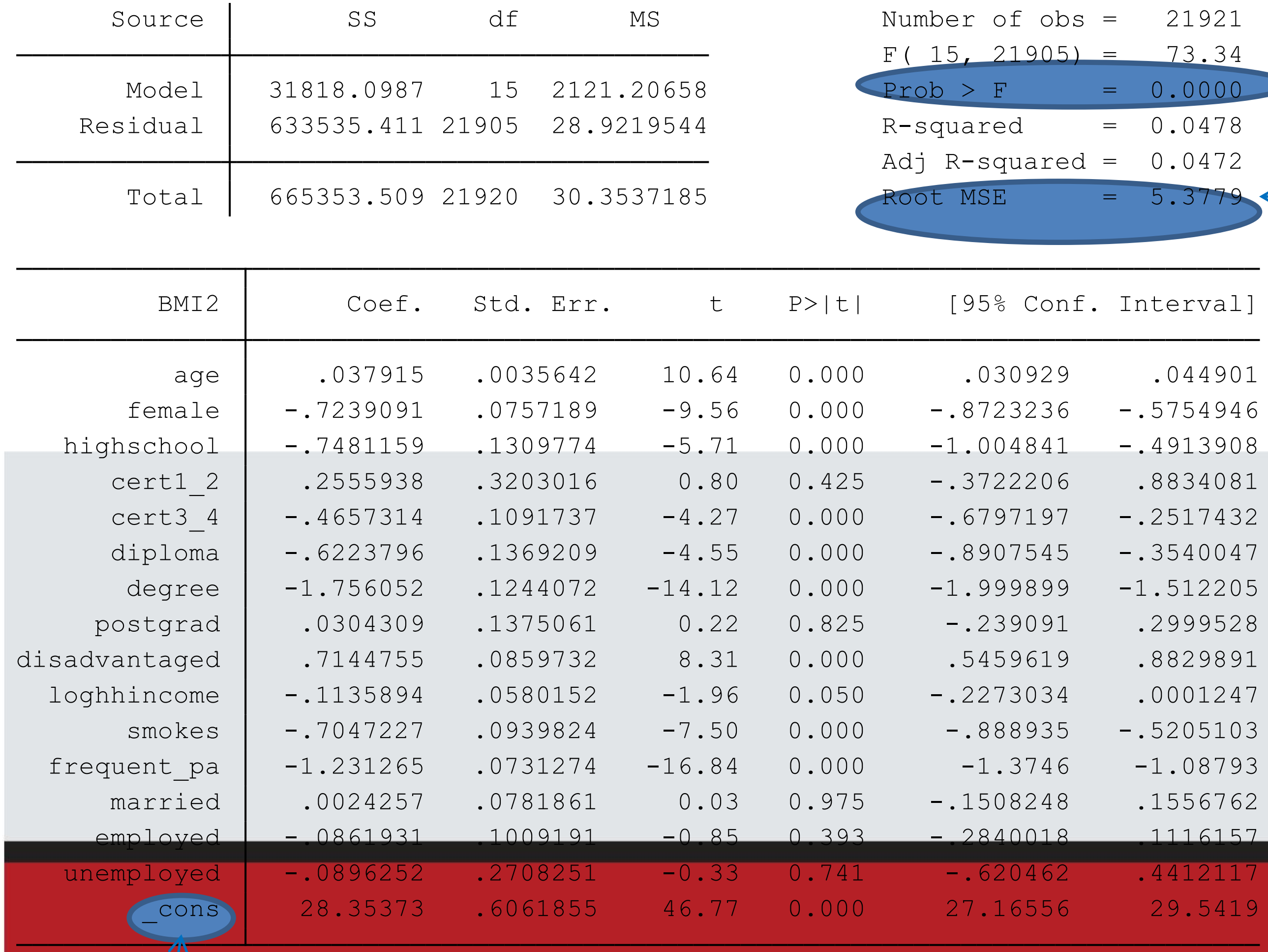
- The model is correctly specified

Data does not violate the assumption of the model you choose

- No exact collinearity between the X variables

Large correlation between two explanatory variables. If this happens you can’t distinguish a separate effect of the variables on the dependent variable

Results (Step 7):



```
      Source |       SS       df       MS              Number of obs =   21921
-------------+------------------------------           F( 15, 21905) =   73.34
       Model |  31818.0987    15  2121.20658           Prob > F      =  0.0000
    Residual |  633535.411 21905  28.9219544           R-squared     =  0.0478
-------------+------------------------------           Adj R-squared =  0.0472
       Total |  665353.509 21920  30.3537185           Root MSE      =  5.3779

------------------------------------------------------------------------------
        BMI2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .037915   .0035642    10.64   0.000      .030929     .044901
      female |  -.7239091   .0757189    -9.56   0.000    -.8723236   -.5754946
  highschool |  -.7481159   .1309774    -5.71   0.000    -1.004841   -.4913908
     cert1_2 |   .2555938   .3203016     0.80   0.425    -.3722206    .8834081
     cert3_4 |  -.4657314   .1091737    -4.27   0.000    -.6797197   -.2517432
     diploma |  -.6223796   .1369209    -4.55   0.000    -.8907545   -.3540047
      degree |  -1.756052   .1244072   -14.12   0.000    -1.999899   -1.512205
    postgrad |   .0304309   .1375061     0.22   0.825     -.239091    .2999528
disadvantaged|   .7144755   .0859732     8.31   0.000     .5459619    .8829891
  loghhincome|  -.1135894   .0580152    -1.96   0.050    -.2273034    .0001247
      smokes |  -.7047227   .0939824    -7.50   0.000     -.888935   -.5205103
 frequent_pa |  -1.231265   .0731274   -16.84   0.000      -1.3746    -1.08793
     married |   .0024257   .0781861     0.03   0.975    -.1508248    .1556762
    employed |  -.0861931   .1009191    -0.85   0.393    -.2840018    .1116157
  unemployed |  -.0896252   .2708251    -0.33   0.741     -.620462    .4412117
       _cons |   28.35373   .6061855    46.77   0.000     27.16556     29.5419
------------------------------------------------------------------------------
```

P-value for whole equation

Square root of residual of model (633535.411) divided by the degrees of freedom (15)

α (the constant term)

Testing for Homoskedasticity (Step 8)

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: age female highschool cert1_2 cert3_4 diploma degree postgrad
                    disadvantaged loghhincome smokes frequent_pa married employed unemployed

         chi2(15)      =  1660.18
         Prob > chi2   =   0.0000
```

- Reject null hypothesis of homoskedasticity
- OLS is not the most efficient model estimate
- Estimated standard errors are incorrect
- F-test is incorrect

Generalised Least Square (Step 6)

- When heteroskedasticity is present, generalised least squares will be a more efficient estimator than ordinary least squares.

- The variance is re-written as: $\mathrm{var}(\varepsilon_i) = \sigma_\alpha^2 + \sigma_\varepsilon^2$

- This is expressed in the error term of our BMI equation:

$$BMI_{it} = \alpha + \beta_1 Individual_{it} + \beta_2 Household_{it} + \beta_3 Health_{it} + \beta_4 Employment + \varepsilon_{it}$$

$$\varepsilon_{it} = \alpha_i + u_{it}$$

Results:

```
Random-effects GLS regression                   Number of obs      =     21921
Group variable: pid                             Number of groups   =      6583

R-sq:  within  = 0.0071                         Obs per group: min =         1
       between = 0.0396                                        avg =       3.3
       overall = 0.0388                                        max =         4

                                                Wald chi2(15)      =    382.16
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        BMI2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0540196   .0053992    10.01   0.000     .0434374    .0646018
      female |  -.5722987   .1286201    -4.45   0.000    -.8243895   -.320208
  highschool |  -.6450872   .2043141    -3.16   0.002    -1.045536   -.2446388
     cert1_2 |  -.1050275   .3927257    -0.27   0.789    -.8747557    .6647007
     cert3_4 |   -.226809   .1601126    -1.42   0.157    -.5406239     .087006
     diploma |  -.6483214   .2162746    -3.00   0.003    -1.072212   -.2244309
      degree |  -1.508137   .1897108    -7.95   0.000    -1.879963    -1.13631
    postgrad |   .0640852   .1910972     0.34   0.737    -.3104584    .4386289
disadvantaged|   .3279032   .0888517     3.69   0.000     .1537571    .5020492
  loghhincome|  -.0030042   .0446496    -0.07   0.946    -.0905158    .0845075
      smokes |   -.371592   .0891074    -4.17   0.000    -.5462392   -.1969448
 frequent_pa |  -.4029532   .0448181    -8.99   0.000    -.4907952   -.3151113
     married |    .151293   .0878546     1.72   0.085    -.0208989    .3234849
    employed |  -.0740042   .0781782    -0.95   0.344    -.2272307    .0792224
  unemployed |  -.1935166   .1451594    -1.33   0.182    -.4780238    .0909907
       _cons |   25.77389   .5304641    48.59   0.000      24.7342    26.81358
-------------+----------------------------------------------------------------
     sigma_u |  4.9345746
     sigma_e |    2.15289
         rho |  .84009172   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Still assume explanatory variables are independent from the error term

To rescale chi stat to F-stat rescale by degrees of freedom: 382.16/15=25.48.

(Error term from this year is correlated with error term from last year)

Inter-class correlation allows for serial correlation in error term

Do we have our best model?

- Two restrictions for ordinary least squares were relaxed in the generalised least square model.

1. Homoskedasticity
2. Serial Correlation

- Still one important assumption which may be violated:
  - Explanatory variables are not correlated with the error term.
  - Not likely to be true.

Endogeneity

- Can lead to bias in the magnitude and significance of your estimated coefficients.
- Three main causes:

1. Direction of relationship does Y cause X or X cause Y?
2. Correlation of explanatory variables with the error term.
3. Omitted variable bias

   Model is missing important variables for explaining the dependent variable

What next then?

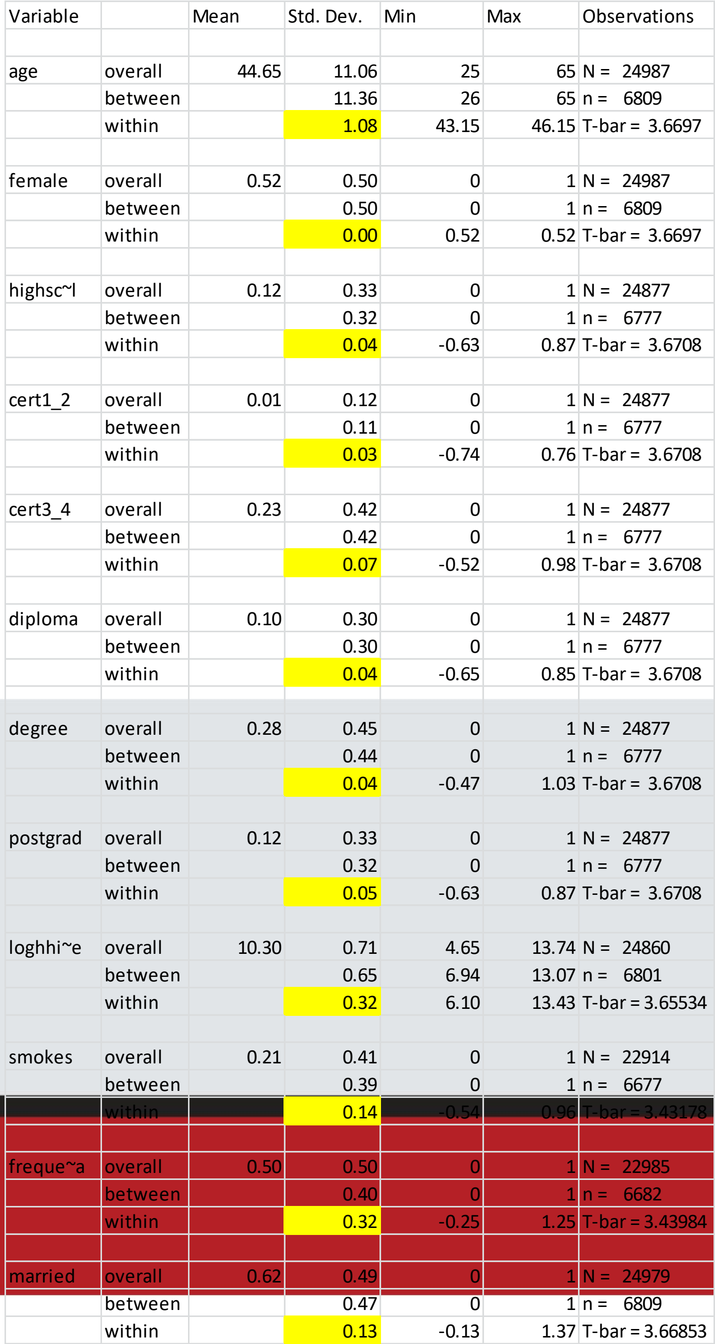
- Fixed effects models removes the bias from correlation of time constant unobserved characteristics.
- Captured by the term, α_i from the error term [redacted] which was modified to control for heteroskedacitiy.
- This bias is removed by effectively taking the mean of all time varying explanatory variables.
- If a variable does not vary over time such as gender it is dropped from the model as the mean would be equal to zero.

Results:

Because mean of the female variable is zero

```
note: female omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =     21921
Group variable: pid                             Number of groups   =      6583

R-sq:  within  = 0.0131                         Obs per group: min =         1
       between = 0.0122                                        avg =       3.3
       overall = 0.0125                                        max =         4

                                                F(14,15324)        =     14.57
corr(u_i, Xb)  = -0.2315                        Prob > F           =    0.0000

------------------------------------------------------------------------------
        BMI2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .1612163   .0141503    11.39   0.000     .1334801    .1889525
      female |          0  (omitted)
  highschool |  -.2996822   .4873314    -0.61   0.539     -1.25491    .6555452
     cert1_2 |  -.6568301   .5701257    -1.15   0.249    -1.774344     .460684
     cert3_4 |   .1238858   .3332258     0.37   0.710    -.5292763     .777048
     diploma |  -.7507973   .5368784    -1.40   0.162    -1.803143    .3015482
      degree |   -.139501   .5260689    -0.27   0.791    -1.170658    .8916565
    postgrad |   .2201849   .3024856     0.73   0.467    -.3727229    .8130927
disadvantaged |    .07102    .108546     0.65   0.513     -.141743    .2837831
  loghhincome |  -.0134885   .0494524    -0.27   0.785    -.1104211    .0834441
      smokes |  -.2607477   .1046585    -2.49   0.013    -.4658909   -.0556045
 frequent_pa |  -.2834061   .0465141    -6.09   0.000    -.3745792    -.192233
     married |   .2074695    .114112     1.82   0.069    -.0162036    .4311426
    employed |  -.0673977   .0854415    -0.79   0.430    -.2348732    .1000777
  unemployed |  -.2287467   .1492554    -1.53   0.125     -.521305    .0638115
       _cons |   20.21425   .7735044    26.13   0.000     18.69809    21.73041
-------------+----------------------------------------------------------------
     sigma_u |  5.3729669
     sigma_e |    2.15289
         rho |  .86165903   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(6582, 15324) =    18.53         Prob > F = 0.0000
```

Looking at variations within variables



Variable		Mean	Std. Dev.	Min	Max	Observations
age	overall	44.65	11.06	25	65	N = 24987
	between		11.36	26	65	n = 6809
	within		1.08	43.15	46.15	T-bar = 3.6697
female	overall	0.52	0.50	0	1	N = 24987
	between		0.50	0	1	n = 6809
	within		0.00	0.52	0.52	T-bar = 3.6697
highsc~l	overall	0.12	0.33	0	1	N = 24877
	between		0.32	0	1	n = 6777
	within		0.04	-0.63	0.87	T-bar = 3.6708
cert1_2	overall	0.01	0.12	0	1	N = 24877
	between		0.11	0	1	n = 6777
	within		0.03	-0.74	0.76	T-bar = 3.6708
cert3_4	overall	0.23	0.42	0	1	N = 24877
	between		0.42	0	1	n = 6777
	within		0.07	-0.52	0.98	T-bar = 3.6708
diploma	overall	0.10	0.30	0	1	N = 24877
	between		0.30	0	1	n = 6777
	within		0.04	-0.65	0.85	T-bar = 3.6708
degree	overall	0.28	0.45	0	1	N = 24877
	between		0.44	0	1	n = 6777
	within		0.04	-0.47	1.03	T-bar = 3.6708
postgrad	overall	0.12	0.33	0	1	N = 24877
	between		0.32	0	1	n = 6777
	within		0.05	-0.63	0.87	T-bar = 3.6708
loghhi~e	overall	10.30	0.71	4.65	13.74	N = 24860
	between		0.65	6.94	13.07	n = 6801
	within		0.32	6.10	13.43	T-bar = 3.65534
smokes	overall	0.21	0.41	0	1	N = 22914
	between		0.39	0	1	n = 6677
	within		0.14	-0.54	0.96	T-bar = 3.43178
freque~a	overall	0.50	0.50	0	1	N = 22985
	between		0.40	0	1	n = 6682
	within		0.32	-0.25	1.25	T-bar = 3.43984
married	overall	0.62	0.49	0	1	N = 24979
	between		0.47	0	1	n = 6809
	within		0.13	-0.13	1.37	T-bar = 3.66853

What next?

- Because there is not much change in the education variables over the 4 years data we have we can not be confident of our findings on these variables from the fixed effect model.
- We can be sure that we have:

1) Endogeneity bias
2) Heteroskedacitiy
3) Serial Correlation

© Ron Leishman * www.ClipartOf.com/440365

Option

- Instrumental Variable Approach
- Find a third variable that is correlated with education but independent of BMI.

- Estimate a proxy fixed effect model that only takes the mean of time varying variables.

Steps 9 & 10

- Negative and significant effect found between BMI and having a degree which held across most model specifications.
- Before suggest an intervention further work is needed to confirm relationship and understand mechanisms.

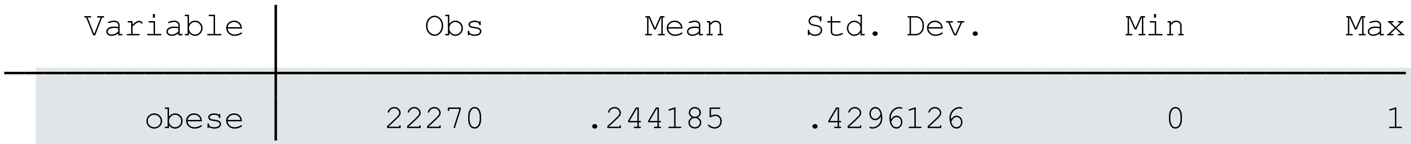
Binary Response Regression Models

- Say you want to narrow the focus of your research question to the determinants of obesity only.

- You take your BMI data and construct a dummy variable for obesity using the WHO classification for obesity.

BMI between 18.5 kg/m^2-24.9 kg/m^2 (healthy weight)
BMI between 25 kg/m^2-29.9 kg/m^2 (overweight)
BMI 30kg/m2 or greater (obese)

Summary of Obesity Variable



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       obese |      22270     .244185    .4296126          0          1
```

Binary Response Regression Models

$$obese_{it} = \begin{cases} 1 if\ BMI \geq 30kg/m^2 \\ 0\ otherwise \end{cases}$$

- $obese_{it}$ can take the values of one with the probability, π_i and zero with the probability $1-\pi_i$.
- The expected mean and variance of $obese_{it}$ will depend upon the underlying probability, π_i

$$E(Obese_i) = \mu_i = \pi_i$$

$$Var(Obese_i) = \sigma_i^2 = \pi_i(1-\pi_i)$$

Binary Response Regression Models

- We violate a main assumption of linear models that explanatory variables can affect the mean but the variance is constant.
- We also need to control for the fact that the dependent variable is truncated between 0 and 1.

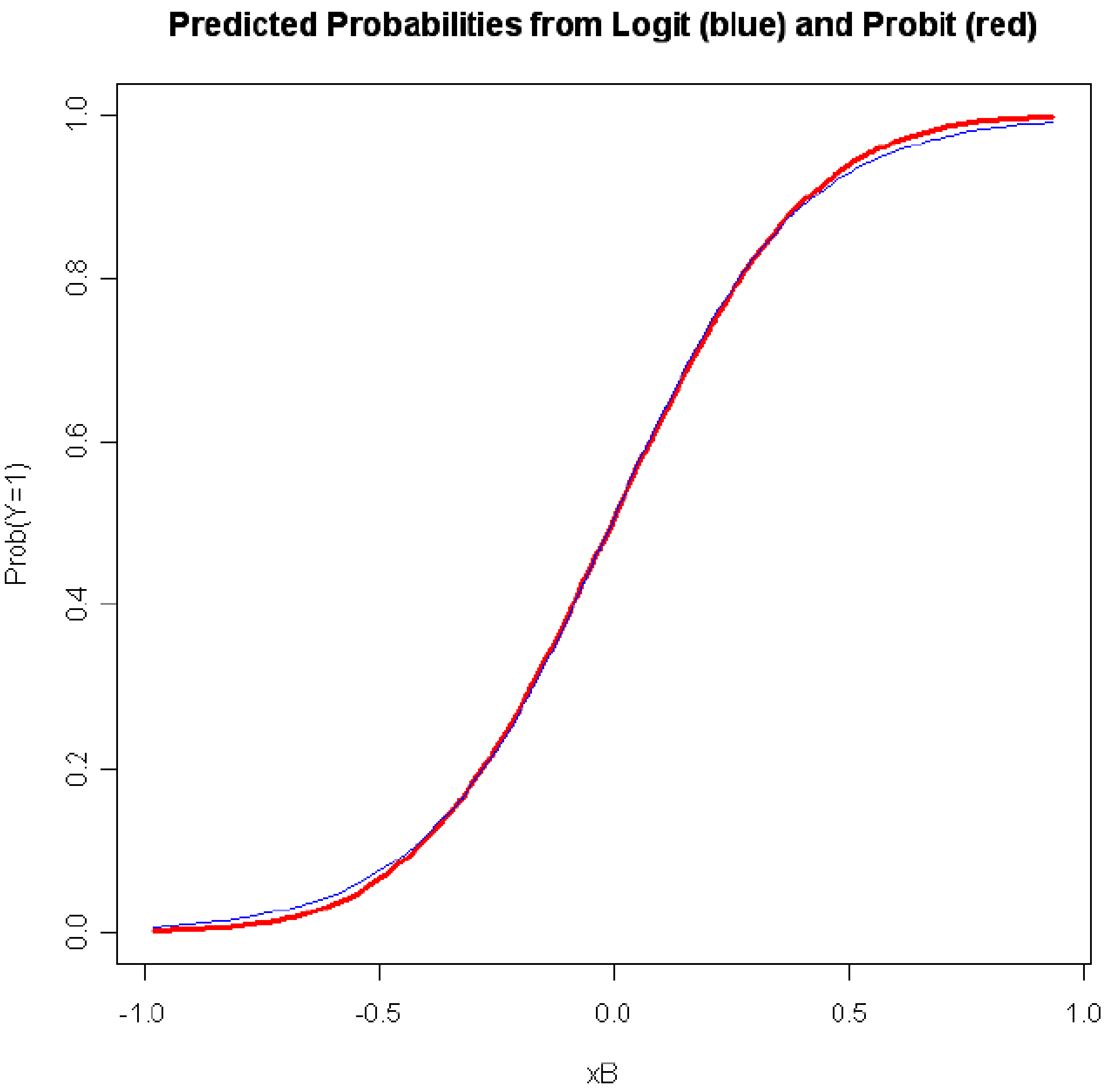
- We need a different type of model:

Two most popular options are:

1. Probit
2. Logit

Probit vs. Logit

- Probit assumes a cumulative standard normal distribution function
- Logit assumes a cumulative logistical function.
- No statistical theory for preferring one over the other
- Results should be similar in a large sample


Predicted Probabilities from Logit (blue) and Probit (red)
Prob(Y=1)
0.0
0.2
0.4
0.6
0.8
1.0
-1.0
-0.5
0.0
0.5
1.0
xB

Probit vs. Logit

- The coefficients from the two models are not directly comparable because they are scaled differently
- Signs and significance will be identical
- The probabilities are virtually the same
- Logit model has fatter tails

Probit Example

Shows overall significance of the model

```
Random-effects probit regression               Number of obs      =     21921
Group variable: pid                            Number of groups   =      6583

Random effects u_i ~ Gaussian                  Obs per group: min =         1
                                                              avg =       3.3
                                                              max =         4

                                               Wald chi2(15)      =    313.64
Log likelihood  = -7268.9695                   Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0232403    .003872     6.00   0.000     .0156514    .0308293
      female |   1.45e-06   .0861945     0.00   1.000    -.1689367    .1689396
  highschool |  -.6569157   .1461271    -4.50   0.000    -.9433195   -.3705119
     cert1_2 |  -.1855968    .329245    -0.56   0.573    -.8309052    .4597116
     cert3_4 |  -.3244582   .1215997    -2.67   0.008    -.5627892   -.0861272
     diploma |  -.6828981   .1535792    -4.45   0.000    -.9839077   -.3818884
      degree |  -1.413866   .1394861   -10.14   0.000    -1.687253   -1.140478
    postgrad |   .3196227   .1462485     2.19   0.029      .032981    .6062644
disadvantaged|   .3416478   .0835444     4.09   0.000     .1779037    .5053919
  loghhincome|  -.0504026   .0461164    -1.09   0.274    -.1407891    .0399839
      smokes |  -.3971027   .0872327    -4.55   0.000    -.5680756   -.2261297
  frequent_pa|  -.4156364   .0506979    -8.20   0.000    -.5150024   -.3162703
     married |   .0042537   .0783037     0.05   0.957    -.1492187    .1577261
    employed |  -.2607877   .0833602    -3.13   0.002    -.4241707   -.0974046
  unemployed |  -.0058434   .1611285    -0.04   0.971    -.3216495    .3099627
       _cons |  -2.783088   .5078715    -5.48   0.000    -3.778497   -1.787678
-------------+----------------------------------------------------------------
    /lnsig2u |   2.845583   .0411215                      2.764986     2.92618
-------------+----------------------------------------------------------------
     sigma_u |   4.148685   .0853001                      3.984824    4.319285
         rho |   .9450899    .002134                      .9407542    .9491255
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =  9004.64 Prob >= chibar2 = 0.000
```

Panel level variance

Standard Deviation

Test of if should control for α_i (random effects)

Average Marginal Effects

- Estimated by calculating individual marginal effects-likelihood of moving from not obese to obese for a one unit change in the explanatory variable in question (estimated for all explanatory variables in the model:

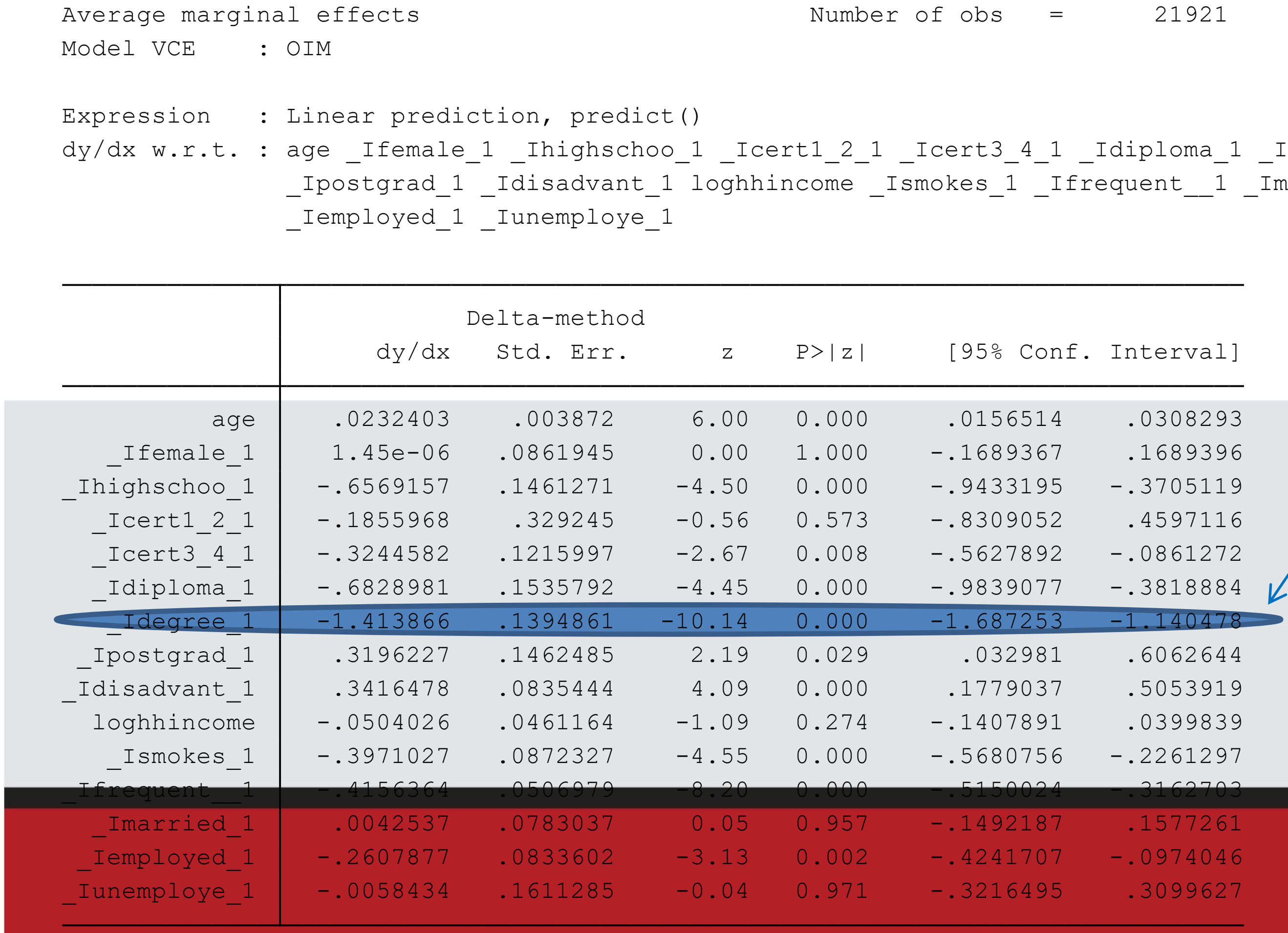
  $(\partial obese_{it} / \partial X_{it})$

- To get average marginal effects, individual marginal effects for all respondents in the sample are averaged.
- This shows the average likelihood of being obese for each explanatory variable

Average Marginal Effects

- For dummy variables, the average marginal effects are calculated by predicting the probability that the dummy variable in question is equal to one and the probability that the dummy variable is equal to zero. The difference between these two probabilities is then averaged across the whole sample.
- For continuous variables, the average marginal effects are estimated by taking the derivative of the predicted probability of the variable in question and averaging across the whole sample.

Average Marginal Effects



```
Average marginal effects                          Number of obs   =      21921
Model VCE    : OIM

Expression   : Linear prediction, predict()
dy/dx w.r.t. : age _Ifemale_1 _Ihighschoo_1 _Icert1_2_1 _Icert3_4_1 _Idiploma_1 _Idegree_1
               _Ipostgrad_1 _Idisadvant_1 loghhincome _Ismokes_1 _Ifrequent__1 _Imarried_1
               _Iemployed_1 _Iunemploye_1

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0232403    .003872     6.00   0.000     .0156514    .0308293
   _Ifemale_1|   1.45e-06   .0861945     0.00   1.000    -.1689367    .1689396
_Ihighschoo_1|  -.6569157   .1461271    -4.50   0.000    -.9433195   -.3705119
  _Icert1_2_1|  -.1855968    .329245    -0.56   0.573    -.8309052    .4597116
  _Icert3_4_1|  -.3244582   .1215997    -2.67   0.008    -.5627892   -.0861272
  _Idiploma_1|  -.6828981   .1535792    -4.45   0.000    -.9839077   -.3818884
   _Idegree_1|  -1.413866   .1394861   -10.14   0.000    -1.687253   -1.140478
  _Ipostgrad_1|   .3196227   .1462485     2.19   0.029      .032981    .6062644
_Idisadvant_1|   .3416478   .0835444     4.09   0.000     .1779037    .5053919
  loghhincome|  -.0504026   .0461164    -1.09   0.274    -.1407891    .0399839
   _Ismokes_1|  -.3971027   .0872327    -4.55   0.000    -.5680756   -.2261297
_Ifrequent__1|  -.4156364   .0506979    -8.20   0.000    -.5150024   -.3162703
  _Imarried_1|   .0042537   .0783037     0.05   0.957    -.1492187    .1577261
 _Iemployed_1|  -.2607877   .0833602    -3.13   0.002    -.4241707   -.0974046
_Iunemploye_1|  -.0058434   .1611285    -0.04   0.971    -.3216495    .3099627
------------------------------------------------------------------------------
```

Problem with model.
Marginal effects shouldn't be greater than 1.
Most likely endogeneity problem.

Logit Example

```
Random effects u_i ~ Gaussian                        Obs per group: min =          1
                                                                    avg =        3.3
                                                                    max =          4

                                                     Wald chi2(15)      =     273.34
Log likelihood  =  -7264.386                         Prob > chi2        =     0.0000

------------------------------------------------------------------------------------
       obese |      Coef.   Std. Err.      z    P>|z|      [95% Conf. Interval]
-------------+----------------------------------------------------------------------
         age |   .0413757   .0073358     5.64   0.000      .0269977    .0557536
      female |  -.0079491   .1685725    -0.05   0.962     -.3383451    .3224469
  highschool |  -1.164756   .2831441    -4.11   0.000     -1.719708   -.6098041
     cert1_2 |  -.3648273   .6302735    -0.58   0.563     -1.600141     .870486
     cert3_4 |  -.5707002   .2342793    -2.44   0.015     -1.029879   -.1115211
     diploma |  -1.205491   .3012874    -4.00   0.000     -1.796004   -.6149789
      degree |  -2.533389   .2708517    -9.35   0.000     -3.064249   -2.002529
    postgrad |   .6075568   .2832631     2.14   0.032      .0523713    1.162742
disadvantaged|   .6053279   .1583307     3.82   0.000      .2950054    .9156504
  loghhincome|  -.0825726   .0848651    -0.97   0.331     -.2489051    .0837599
      smokes |  -.7515844   .1653225    -4.55   0.000     -1.075611   -.4275582
 frequent_pa |   -.755414   .0928977    -8.13   0.000     -.9374901   -.5733378
     married |   .0193703   .1489499     0.13   0.897     -.2725661    .3113067
    employed |  -.4732393   .1533456    -3.09   0.002     -.7737911   -.1726876
  unemployed |  -.0192225   .2899264    -0.07   0.947     -.5874678    .5490227
       _cons |  -5.162994   .9423629    -5.48   0.000     -7.009992   -3.315997
-------------+----------------------------------------------------------------------
    /lnsig2u |    4.05162   .0420026                       3.969297    4.133944
-------------+----------------------------------------------------------------------
     sigma_u |   7.582251   .1592371                       7.276488    7.900862
         rho |   .9458729   .0021504                       .9415001    .9499362
------------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =  9011.46 Prob >= chibar2 = 0.000
```

Logit Example (Odds ratio)

Odds-ratio

```
Random-effects logistic regression              Number of obs      =     21921
Group variable: pid                             Number of groups   =      6583

Random effects u_i ~ Gaussian                   Obs per group: min =         1
                                                               avg =       3.3
                                                               max =         4

                                                Wald chi2(15)      =    273.34
Log likelihood  =  -7264.386                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       obese |         OR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.042244   .0076457     5.64   0.000     1.027365    1.057337
      female |   .9920824   .1672378    -0.05   0.962     .7129492    1.380502
  highschool |   .3119987   .0883406    -4.11   0.000     .1791184    .5434573
     cert1_2 |   .6943166   .4376093    -0.58   0.563     .2018681    2.388071
     cert3_4 |   .5651296   .1323982    -2.44   0.015     .3570501    .8944725
     diploma |   .2995448   .0902491    -4.00   0.000     .1659608    .5406523
      degree |   .0793895   .0215028    -9.35   0.000     .0466889    .1349934
    postgrad |    1.83594   .5200542     2.14   0.032     1.053767    3.198693
disadvantaged|   1.831853   .2900386     3.82   0.000     1.343134      2.4984
  loghhincome |   .9207446   .0781391    -0.97   0.331      .779654    1.087368
      smokes |   .4716187   .0779692    -4.55   0.000     .3410894    .6520994
  frequent_pa |   .4698161   .0436448    -8.13   0.000     .3916095     .563641
     married |   1.019559   .1518632     0.13   0.897     .7614231    1.365208
    employed |   .6229809   .0955314    -3.09   0.002     .4612611    .8414004
  unemployed |    .980961   .2844065    -0.07   0.947     .5557327     1.73156
       _cons |   .0057245   .0053946    -5.48   0.000     .0009028    .0362978
-------------+----------------------------------------------------------------
    /lnsig2u |    4.05162   .0420026                      3.969297    4.133944
-------------+----------------------------------------------------------------
     sigma_u |   7.582251   .1592371                      7.276488    7.900862
         rho |   .9458729   .0021504                      .9415001    .9499362
------------------------------------------------------------------------------
Likelihood-ratio test of rho=0: chibar2(01) =  9011.46 Prob >= chibar2 = 0.000
```

Comparing logit and probit coefficients

	Probit	Logit
obese	Coef.	Coef.
age	0.02	0.04
_Ifemale_1	0.00	-0.01
_Ihighschoo_1	-0.66	-1.16
_Icert1_2_1	-0.19	-0.36
_Icert3_4_1	-0.32	-0.57
_Idiploma_1	-0.68	-1.21
_Idegree_1	-1.41	-2.53
_Ipostgrad_1	0.32	0.61
_Idisadvant_1	0.34	0.61
loghhincome	-0.05	-0.08
_Ismokes_1	-0.40	-0.75
_Ifrequent__1	-0.42	-0.76
_Imarried_1	0.00	0.02
_Iemployed_1	-0.26	-0.47
_Iunemploye_1	-0.01	-0.02
_cons	-2.78	-5.16

Decompositons

- If you have potential pathways that you think may explain observed composition
- Three main ways to implement:

1) Oaxaca Blinder
2) KHB
3) nldecompose

Oaxaca method

- Study distributional differences between two groups

$$\begin{aligned}\ln \bar{W}_M - \ln \bar{W}_N = \;& \widehat{\kappa}_M, \widehat{\gamma}_M, \hat{\xi}_M, \widehat{\psi}_M \left(\bar{X}_M - \bar{X}_N, \bar{H}_M - \bar{H}_N, \bar{L}_M - \bar{L}_N, \bar{B}_M - \bar{B}_N \right) \qquad (5) \\ & + \bar{X}_N, \bar{H}_N, \bar{L}_N, \bar{B}_N \left(\widehat{\kappa}_M - \widehat{\kappa}_N, \widehat{\gamma}_M - \widehat{\gamma}_N, \hat{\xi}_M - \hat{\xi}_N, \widehat{\psi}_M - \widehat{\psi}_N \right),\end{aligned}$$

KHB Method

- Addresses problem caused by the need for rescaling or attenuation bias in non-linear models
- Used to explain how a mediator (pathway) variable Z explains the relationship between X and a latent outcome variable Y

KHB Method

- Over comes the scaling problem by taking out of Z the information that is not in X by calculating the residuals of a linear regression of Z on X.
- R is the used instead of Z in a reduced form model following a similar format to the standard Oaxaca linear decomposition

Nldecompose

- Similar idea to overcome scaling problem for non-linear models
- More flexible than khb method can be used with ordered logit and probit

References

- Sinning, M., Hahn, M., & Bauer, T. K. (2008). The Blinder–Oaxaca decomposition for nonlinear regression models. *The Stata Journal*, *8*(4), 480-492.
- Kohler, U., & Karlson, K. (2010). KHB: Stata module to decompose total effects into direct and indirect via KHB-method.
- Brown, H. (2011). Marriage, BMI, and wages: A double selection approach. *Scottish Journal of Political Economy*, *58*(3), 347-377.