DATA ANALYSIS STRATEGIES

ESTER CERIN

WHAT WILL THIS WORKSHOP BE ABOUT?

Goals

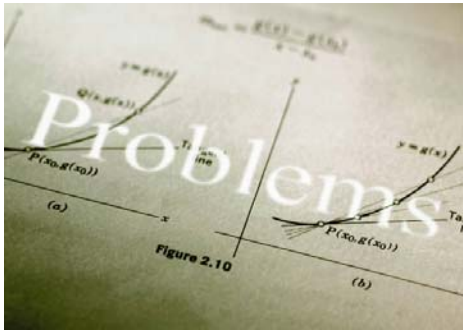
- Identify main challenges encountered in conducting pooled analyses for the IPEN Adult study (multi-site study aiming to examine the strength and shape of environment-physical activity relationships)
- Present analytical approaches that can be applied to any study facing similar complex problems

Learning objectives (learn …)

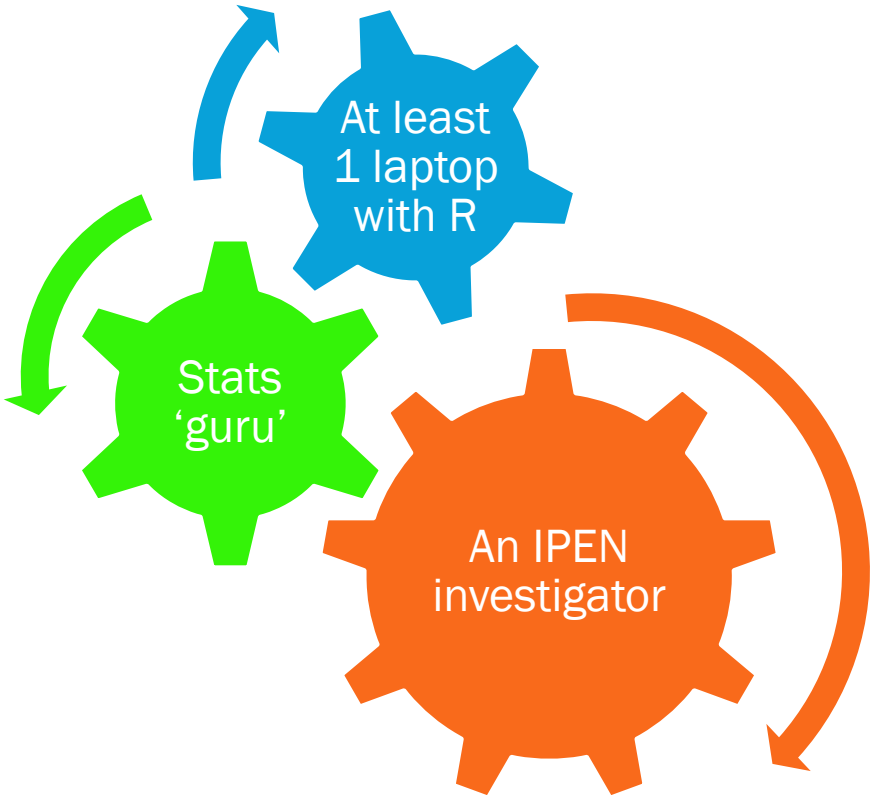
- Fundamentals of models appropriate for correlated, non-normally distributed, non-linearly related variables (Generalized Additive Mixed Models -> GAMMs)
- How to choose the most appropriate non-normal error distribution
- How to interpret results from various GAMMs and report findings in a scientific article
- How to explore associations of environmental characteristics with physical activity at the within-site and between-site levels

OVERVIEW OF TODAY'S WORKSHOP

1. Working groups (4-5 people) (10 min)
2. **Introduction to the IPEN studies:** analytical challenges (15 min)
3. **CHALLENGE 1**: dealing with correlated data (30 min)
4. **CHALLENGE 2**: dealing with non-normally distributed data (35 min)
5. BREAK (15 min)
6. **CHALLENGE 2**: dealing with non-normally distributed data (30 min)
7. **CHALLENGE 3**: dealing with curvilinear relationships (20 min)
8. **CHALLENGE 4** : estimating within- and between-site effects (15 min)
9. Wrapping up (10 min)



WORKING GROUPS (4-5 PEOPLE)



INTRODUCTION TO

Cross-sectional, observational, multi-site study adopting a two-stage stratified sampling strategy

Regression?

IPEN Study Aims

The primary aim of the IPEN study was to estimate strengths of association between detailed measures of the neighborhood built environment with leisure physical activity, walking/cycling for transportation, and BMI in all participants, based on self-report survey data collected according to a common protocol. The secondary aims of the IPEN study examined the same questions as the primary aims, but used objective measures in a smaller sample of participants:

Journal of Physical Activity and Health, 2013, 10, 581-601

JOURNAL OF
Physical Activity & Health
Official Journal of ISPAH
www.JPAH-Journal.com
ORIGINAL RESEARCH

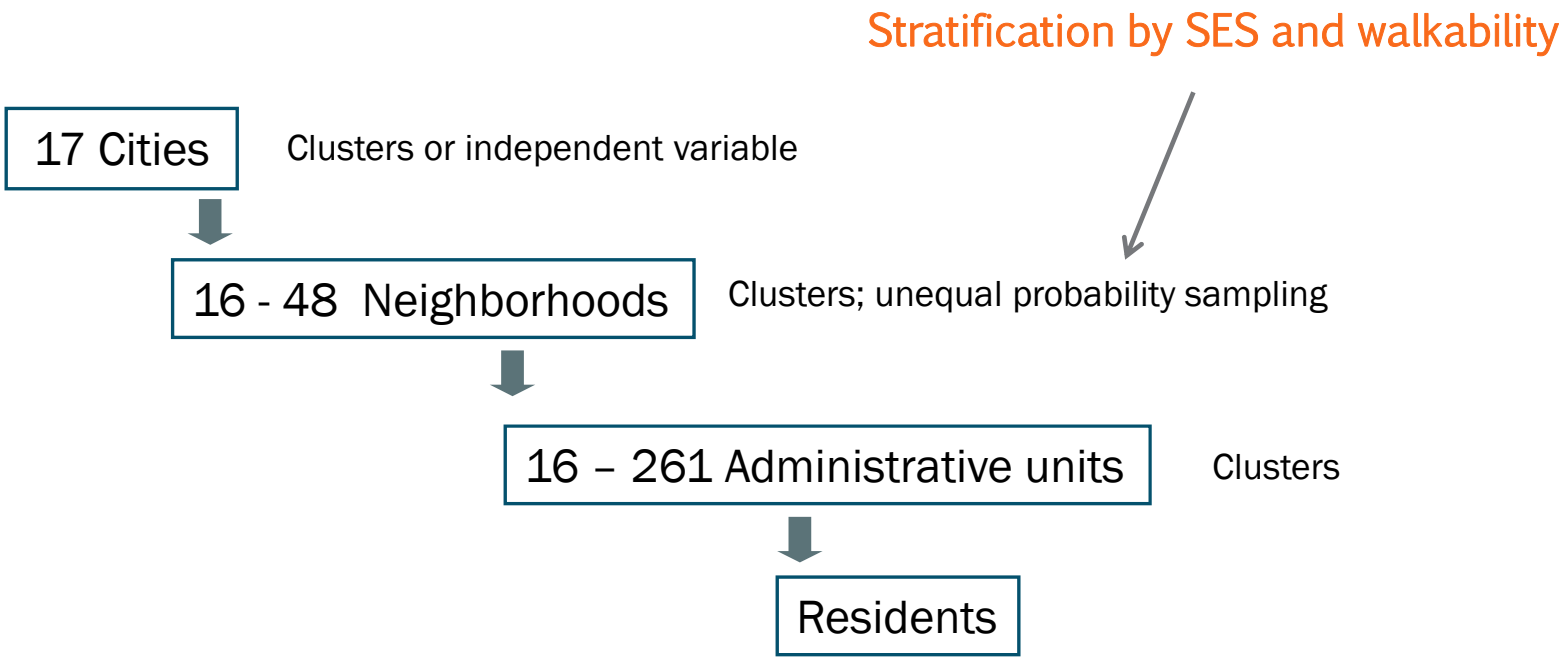
Advancing Science and Policy Through a Coordinated International Study of Physical Activity and Built Environments: IPEN Adult Methods

Jacqueline Kerr, James F. Sallis, Neville Owen, Ilse De Bourdeaudhuij, Ester Cerin, Takemi Sugiyama, Rodrigo Reis, Olga Sarmiento, Karel Frömel, Josef Mitáš, Jens Troelsen, Lars Breum Christiansen, Duncan Macfarlane, Deborah Salvo, Grant Schofield, Hannah Badland, Francisco Guillen-Grima, Ines Aguinaga-Ontoso, Rachel Davey, Adrian Bauman, Brian Saelens, Chris Riddoch, Barbara Ainsworth, Michael Pratt, Tom Schmidt, Lawrence Frank, Marc Adams, Terry Conway, Kelli Cain, Delfien Van Dyck, and Nicole Bracy

INTRODUCTION TO IPEN

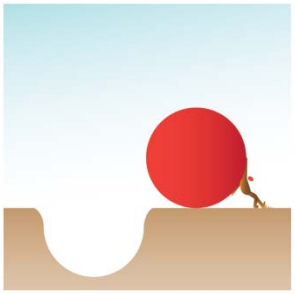
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

Cross-sectional, observational, multi-site study adopting a two-stage stratified sampling strategy



IPEN PROJECT = COMPLEX

CHALLENGE 1



Correlated data … Correlated residuals

Violation of independence assumption

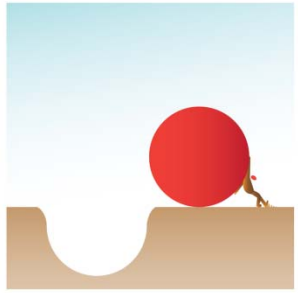
Consequences

- Incorrect standard errors
- Clustering primarily affects variance or precision of estimation rather than bias (unless individual-level associations between factors measured at the individual level differ from those at the area-level)

BAD NEWS: can't use "standard" OLS regression models, generalized linear models, generalized additive models …

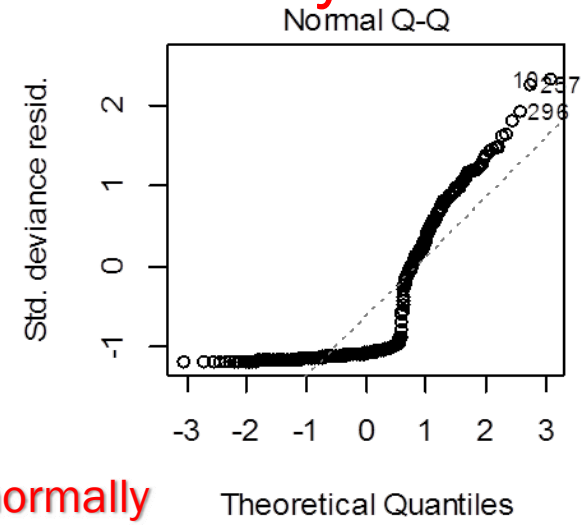
IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

CHALLENGE 2

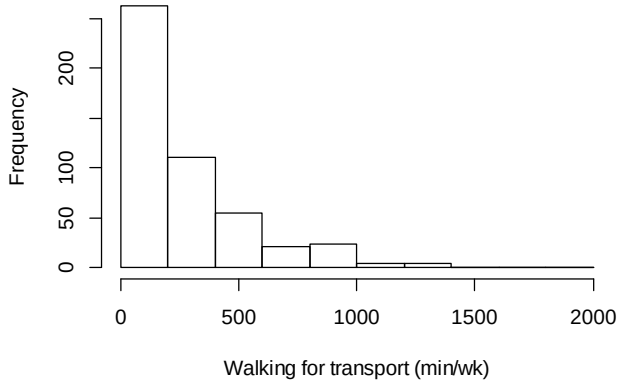


Non-normally distributed data & heteroscedasticity

- Positive skew
- A lot of zero values …
  - Leisure-time physical activity
  - Walking for transport

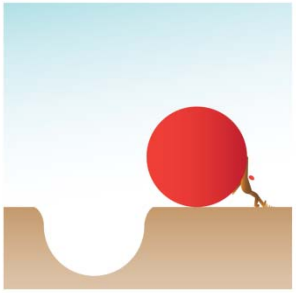


BAD NEWS: can't use regression models assuming normally distributed residuals …

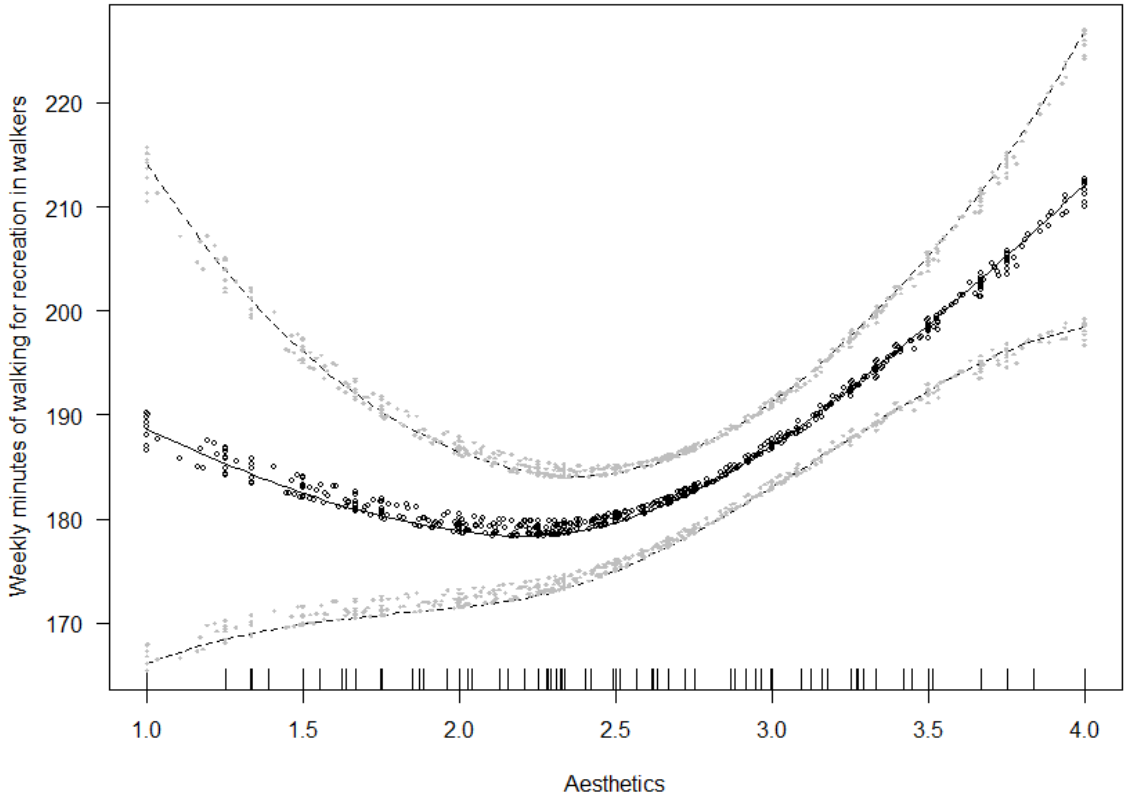


IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

CHALLENGE 3



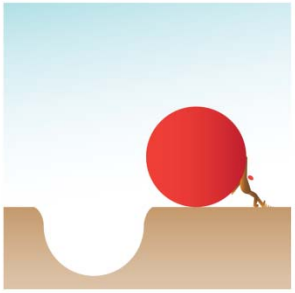
Non-linear relationships



BAD NEWS: can't simply enter 'untransformed' predictors in the regression models …

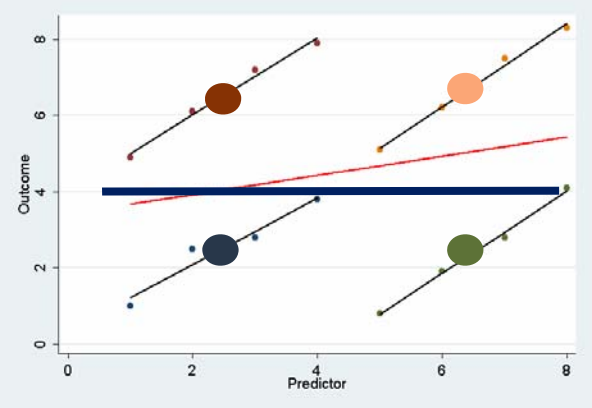
IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

CHALLENGE 4



Between- and within-city associations

- The strength of these may differ



- Need to assess within- and between-city effects

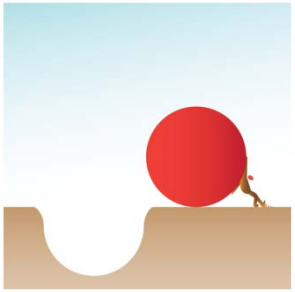
$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 \overline{X}_j$$

Within-city

Between-city

IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

CHALLENGE 4



Between- and within-city associations

- The strength of these may differ
- Need to assess within- and between-city effects

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 \overline{X}_j$$

Within-city

Between-city

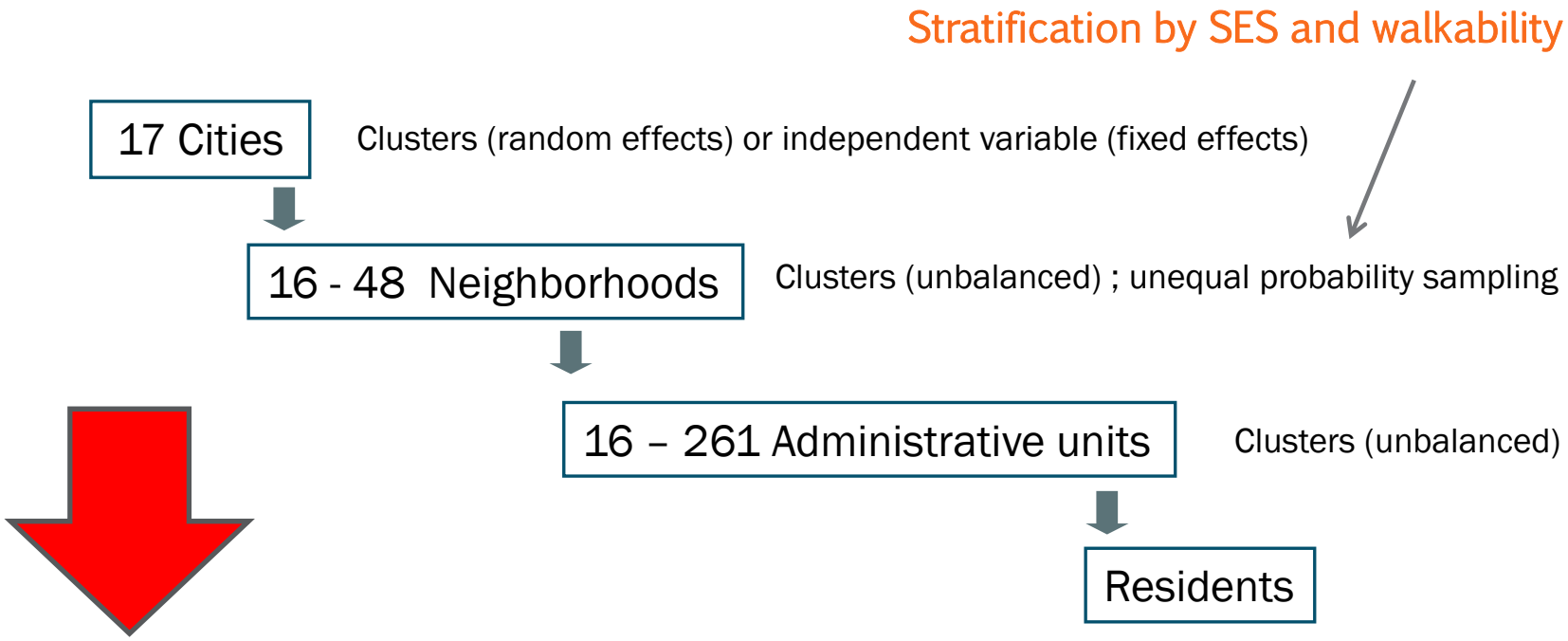
BUT BAD NEWS: we cannot easily distinguish between- and within-city effects …

- IPEN "sampled" only 17 cities
- Cities are not truly "random" factors
- Regression models encompass many predictors

IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

CHALLENGE 1:
DEALING WITH CORRELATED DATA

Cross-sectional, observational, multi-site study adopting a two-stage stratified sampling strategy



Questions:

What sources of dependency do we have? How many?

Do we model these sources of dependency as random or fixed factors?

How do we model unequal probability sampling of neighborhoods?

‘CANDIDATE’ REGRESSION MODELS

1. Generalized linear models with robust standard errors?

   Statistically inefficient

2. Generalized estimating equations?

   Not appropriate with highly unbalanced clusters

3. Linear mixed models?

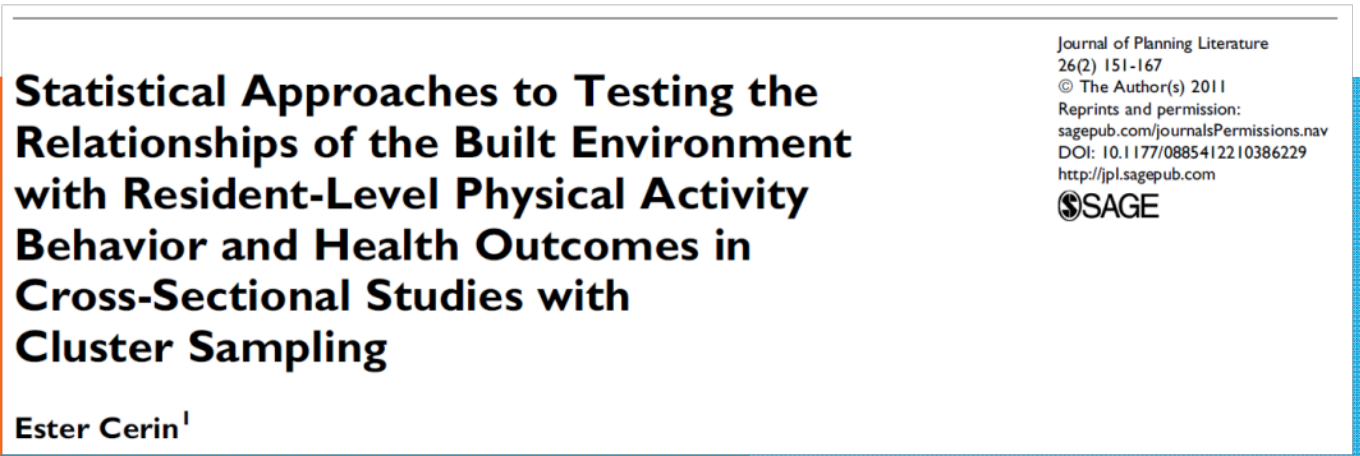
   Yes, if you expect normally-distributed errors and linear associations

4. Generalized linear mixed models (*aka* multilevel generalized linear models)?

   Yes, if you expect linear associations

5. Generalized additive mixed models?

   Yes, if you do not expect any of the above.



Statistical Approaches to Testing the Relationships of the Built Environment with Resident-Level Physical Activity Behavior and Health Outcomes in Cross-Sectional Studies with Cluster Sampling

Ester Cerin[1]

Journal of Planning Literature
26(2) 151-167

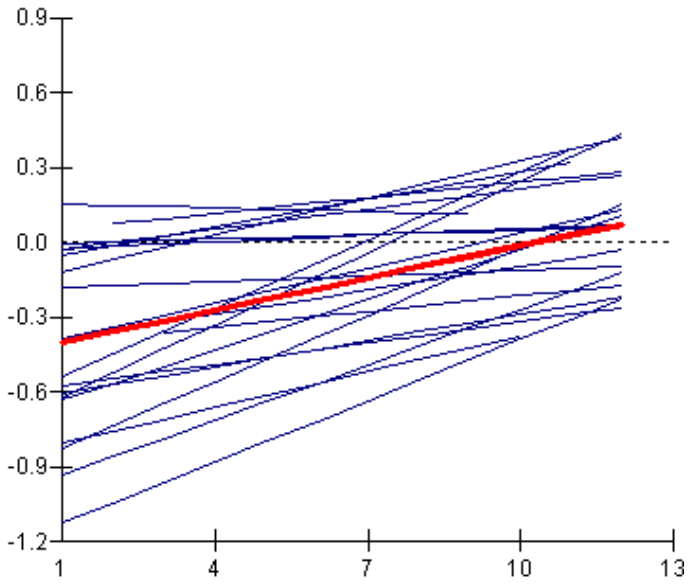
DOI: 10.1177/0885412210386229
http://jpl.sagepub.com
SAGE

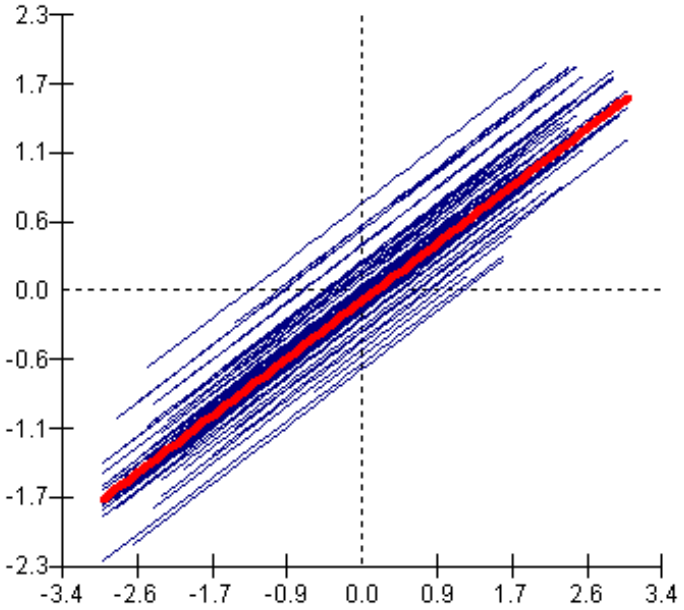
IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

THE 'MIXED' IN GENERALIZED ADDITIVE MIXED MODELS: DEALING WITH CORRELATED DATA

... also referred to as "hierarchical"

... OR "multilevel"

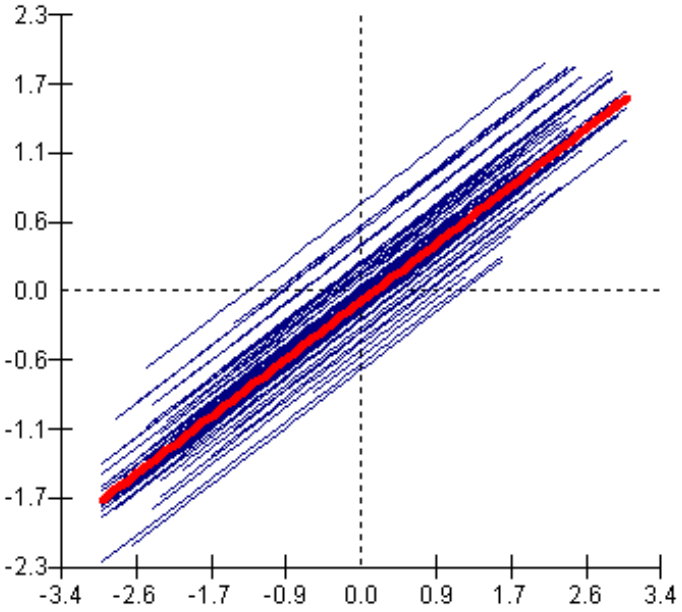




IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

THE 'MIXED' IN GENERALIZED ADDITIVE MIXED MODELS: DEALING WITH CORRELATED DATA

Random intercept model



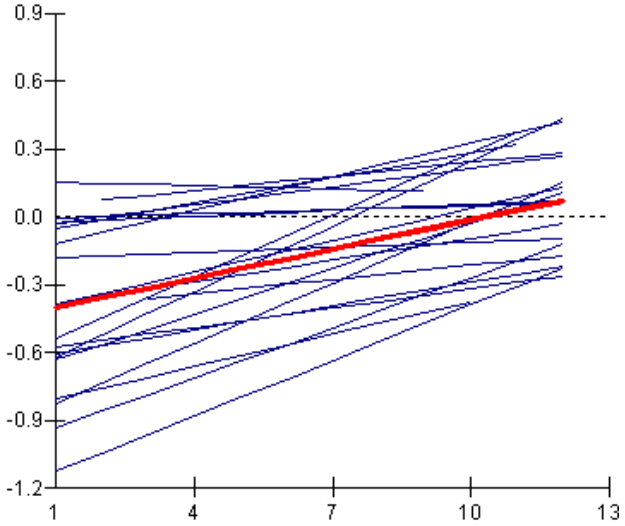
$$Walking_{ij} = \beta_{0j} cons + \beta_1 (R.Density)_{ij} + \beta_2 (SES)_j + \beta_3 (Cities)_j + \beta_4 (\text{cov})_{ij}$$

$$\beta_{0j} = \beta_0 + u_{0j} + e_{0ij}$$

$$u_{0j} \sim N(0, \sigma^2_{u0}) \quad e_{0ij} \sim N(0, \sigma^2_{e0})$$

THE 'MIXED' IN GENERALIZED ADDITIVE MIXED MODELS: DEALING WITH CORRELATED DATA

Random intercept and random slope model



$$Walking_{ij} = \beta_{0ij} cons + \beta_{1j}(R.Density)_{ij} + \beta_2(SES)_j + \beta_3(Cities)_j + \beta_4(\text{cov})_{ij}$$

$$\beta_{0j} = \beta_0 + u_{0j} + e_{0ij}$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

$$u_{0j} \sim N(0, \sigma^2_{u0}) \quad e_{0ij} \sim N(0, \sigma^2_{e0})$$

$$u_{1j} \sim N(0, \sigma^2_{u1}) \quad Cov(u_{0j}, u_{1j}) = \rho\sigma_{u0}\sigma_{u1}$$

GAMMS IN 'R'

RGui (64-bit) - [Data Editor]

File Windows Edit Help

	Part	Country	City	SES_final	Walk_final	Age_final	Gender_final	Marital_2grp_final	Educ_3grp_final	Job_final	i_lei_1	I_LeiWlkPA	GN_ResidDen	GI
139	139	1	16	0	1	38	1	0	2	1	7	280	39	4
140	140	1	16	0	1	36	1	1	3	1	2	120	39	3
141	141	1	16	0	1	42	1	1	3	1	1	30	14	3
142	142	1	16	0	1	52	1	0	3	1	2	40	50	2
143	143	1	16	0	1	32	2	0	3	1	1	30	85	4
144	144	1	16	0	1	52	1	1	3	1	5	225	39	3
145	145	1	16	0	1	37	2	1	2	1	3	360	61	3
146	146	1	16	0	1	45	2	1	3	1	0	0	25	3
147	147	1	16	0	1	35	2	1	3	1	1	20	39	2
148	148	1	16	0	1	53	2	0	2	0	7	420	108	2
149	149	1	16	0	1	31	1	0	3	1	3	180	39	3
150	150	1	16	0	1	31	1	0	2	1	0	0	14	3
151	151	1	16	0	1	50	1	0	2	1	2	20	24	3
152	152	1	16	0	1	35	1	1	3	1	5	100	249	4
153	153	1	16	0	1	53	1	1	3	1	4	180	14	2
154	154	1	16	0	1	40	2	1	2	0	3	90	59	3
155	155	1	16	0	1	37	1	0	3	1	7	315	48	3
156	156	1	16	0	1	53	2	0	3	1	7	140	14	2
157	157	1	16	0	1	28	2	0	3	1	0	0	34	4
158	158	1	16	0	1	63	1	1	2	0	0	0	361	3
159	159	1	16	0	1	35	2	0	3	1	1	30	39	3
160	160	1	16	0	1	28	1	0	2	1	2	480	245	1
161	161	1	16	0	1	35	1	0	3	0	2	120	133	1
162	162	1	16	0	1	37	1	0	3	1	2	120	109	4
163	163	1	16	0	1	44	2	0	3	1	0	0	134	4
164	164	1	16	0	1	30	2	0	3	1	2	80	98	4
165	165	1	16	0	1	34	1	1	2	1	3	90	184	4
166	166	1	16	0	1	35	1	0	3	1	5	225	85	4
167	167	1	16	0	1	36	1	0	3	1	2	30	208	4
168	168	1	16	0	1	49	1	0	3	1	7	630	135	4

Association of perceived residential density with weekly minutes of walking for recreation

GAMMs in R - dealing with correlated data

Setting up GAMMs in 'R' (random intercept)

'mgcv' library

Assuming normally-distributed residuals and linear relationships

	Model 1	Model 2
Outcome	l_LeiWlkPA	l_LeiWlkPA
Predictor	GN_ResidDen	GN_ResidDen
Design factor	fSES	fSES
Covariates	Age_final, fgender, fjob, feducation, fmarital	Age_final, fgender, fjob, feducation, fmarital
Administrative units	Random factor (cluster)	Random factor (cluster)
Cities	Random factor (City)	Fixed factor (fcity)

Model.1<-(gamm(l_LeiWlkPA ~ fSES + fgender + feducation + fjob + fmarital + Age_final + GN_ResidDen, data=complete, **random=list(City=~1, cluster=~1)**))

Model.2<-(gamm(l_LeiWlkPA ~ fSES + fgender + feducation + fjob + fmarital + Age_final + fcity + GN_ResidDen, data=complete, **random=list(cluster=~1)**))

MODEL 1 (lme component)

```
Linear mixed-effects model fit by maximum likelihood
 Data: strip.offset(mf)
     AIC      BIC   logLik
  175745 175834.6 -87860.5

Random effects:
 Formula: ~1 | City
        (Intercept)
StdDev:     55.2626

 Formula: ~1 | cluster %in% City
        (Intercept) Residual
StdDev:    18.39642 216.2723

Fixed effects: y ~ X - 1
                                              Value Std.Error    DF   t-value p-value
X(Intercept)                               61.19865 16.933450 12538  3.614068  0.0003
XfSESHigh                                   0.56017  4.605586 12538  0.121628  0.9032
XfgenderFemale                             -3.53170  3.923731 12538 -0.900086  0.3681
XfeducationHigh school or some college     11.40928  6.122696 12538  1.863440  0.0624
XfeducationCollege or more                  0.69183  6.390216 12538  0.108264  0.9138
Xfjob1                                    -20.40180  4.582352 12538 -4.452256  0.0000
XfmaritalWith partner                      -6.18400  4.084596 12538 -1.513980  0.1301
XAge_final                                  1.48872  0.159747 12538  9.319209  0.0000
XGN_ResidDen                                0.14735  0.026174 12538  5.629545  0.0000
```

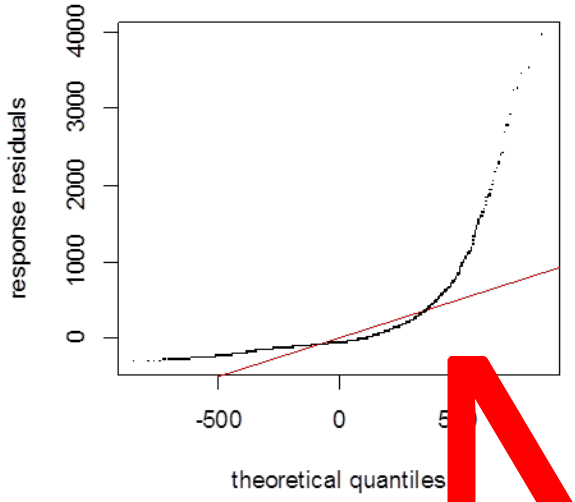
MODEL 1 (gam component)

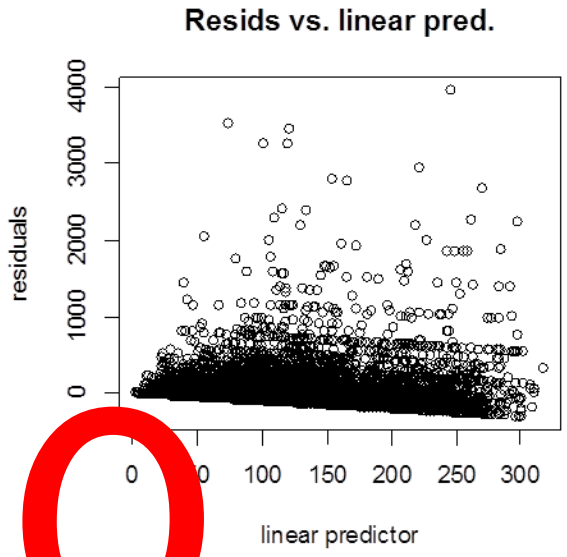
```
Family: gaussian
Link function: identity

Formula:
I_LeiWlkPA ~ fSES + fgender + feducation + fjob + fmarital +
    Age_final + GN_ResidDen

Parametric coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             61.19865   16.93345   3.614 0.000303 ***
fSESHigh                                 0.56017    4.60559   0.122 0.903195
fgenderFemale                           -3.53170    3.92373  -0.900 0.368091
feducationHigh school or some college   11.40928    6.12270   1.863 0.062423 .
feducationCollege or more                0.69183    6.39022   0.108 0.913788
fjob1                                  -20.40180    4.58235  -4.452 8.57e-06 ***
fmaritalWith partner                    -6.18400    4.08460  -1.514 0.130055
Age_final                                1.48872    0.15975   9.319  < 2e-16 ***
GN_ResidDen                              0.14735    0.02617   5.630 1.84e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


R-sq.(adj) =  0.0133  Scale est. = 46774     n = 12919
```

MODEL 2 (gam component)

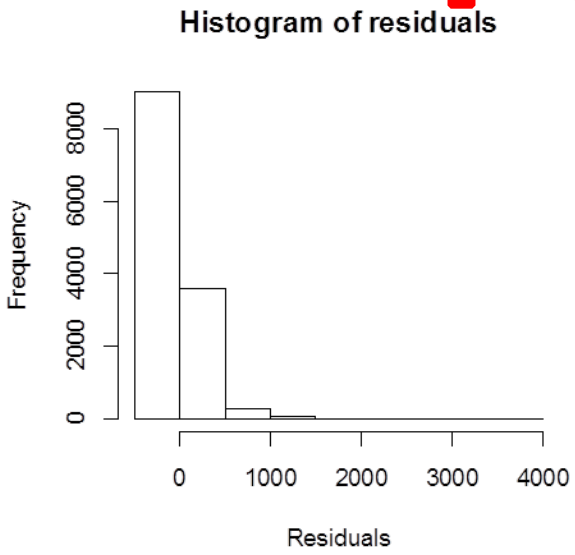
```
Family: gaussian
Link function: identity

Formula:
I_LeiWlkPA ~ fSES + fgender + feducation + fjob + fmarital +
    Age_final + fcity + GN_ResidDen

Parametric coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                             67.70935   11.26873   6.009 1.92e-09 ***
fSESHigh                                 0.42056    4.49791   0.094 0.925506
fgenderFemale                           -3.56512    3.92437  -0.908 0.363655
feducationHigh school or some college   11.19197    6.12782   1.826 0.067810 .
feducationCollege or more                0.58783    6.39424   0.092 0.926754
fjob1                                  -20.40682    4.58240  -4.453 8.53e-06 ***
fmaritalWith partner                    -6.15160    4.08382  -1.506 0.132006
Age_final                                1.49978    0.15971   9.390  < 2e-16 ***
fcityGhent, Belgium                    -46.65139    9.22375  -5.058 4.30e-07 ***
…
fcityPamplona, Spain                    97.99763   10.84807   9.034  < 2e-16 ***
…
fcityBaltimore, USA                    -22.83341    9.98015  -2.288 0.022161 *
GN_ResidDen                              0.14783    0.02639   5.601 2.18e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


R-sq.(adj) =  0.0587  Scale est. = 46763     n = 12919
```

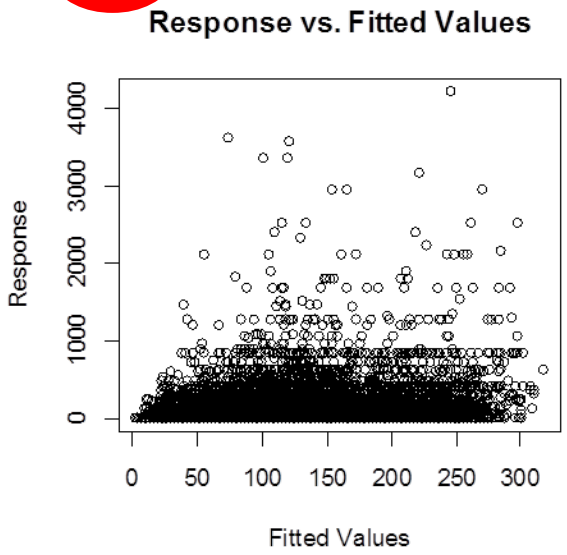
ARE THESE MODELS VALID? DIAGNOSTICS …



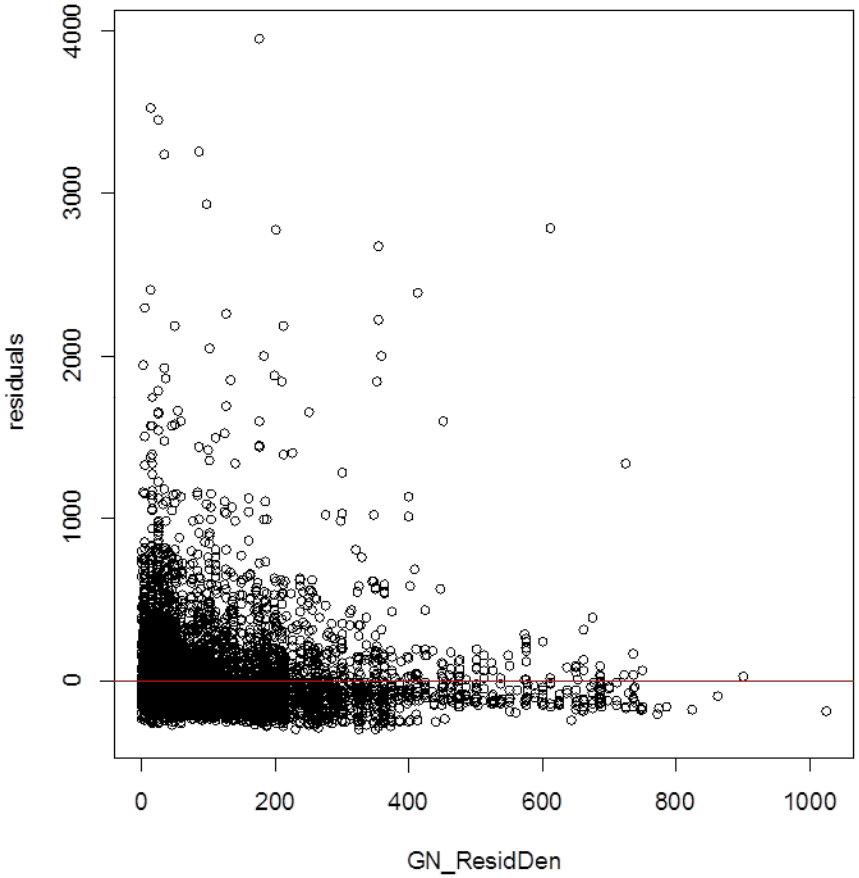


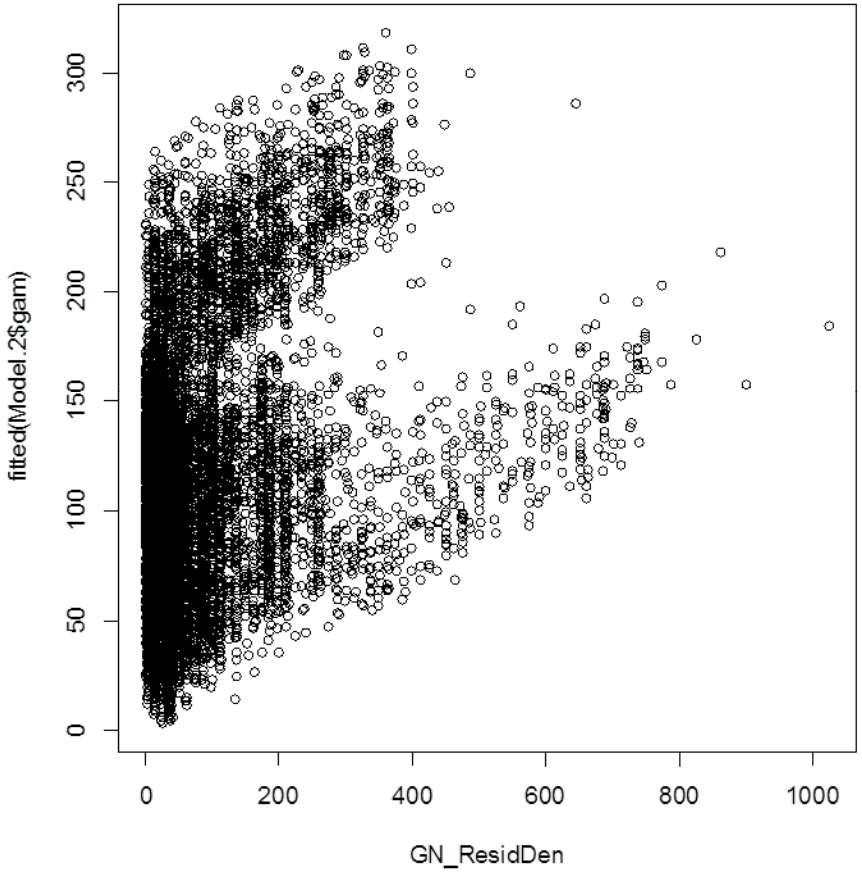
NO



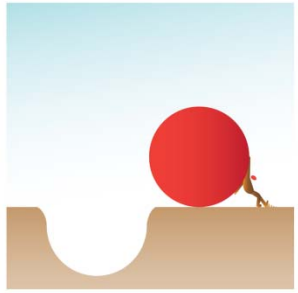


ARE THESE MODELS VALID? DIAGNOSTICS …



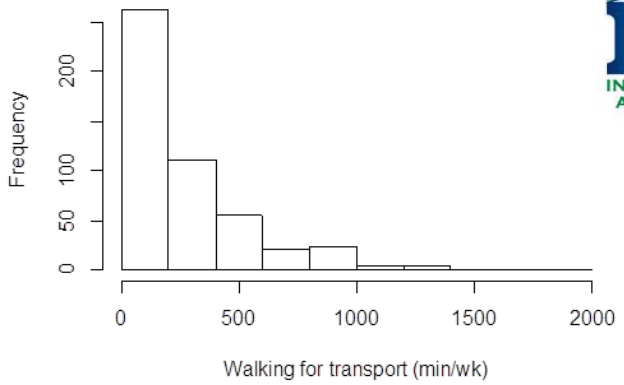


CHALLENGE 2

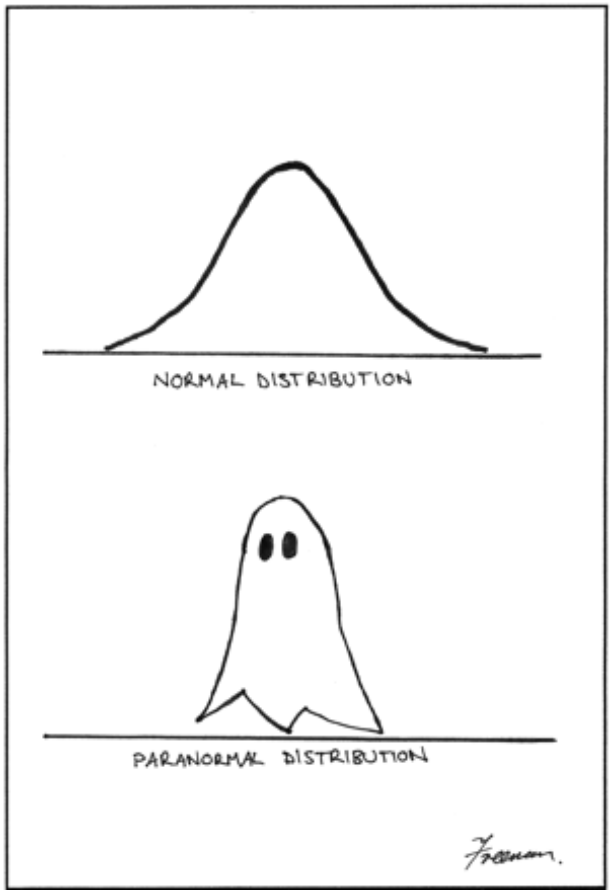


Non-normally distributed data & heteroscedasticity

- Positive skew
- A lot of zero values …
  - Leisure-time physical activity
  - Walking for transport



IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

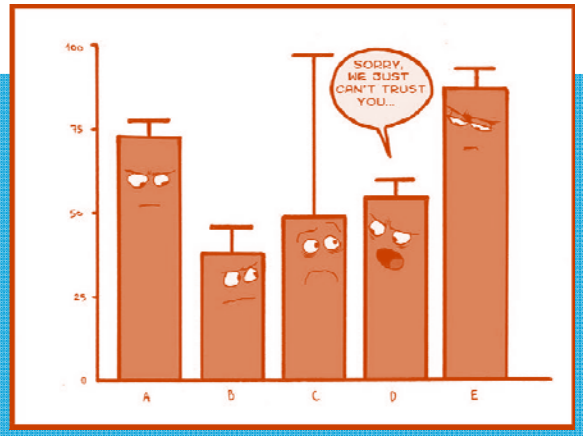


WHAT CAN WE DO?

Box Cox transformation – helps make the data normal

PROBLEMS WITH TRANFORMATIONS

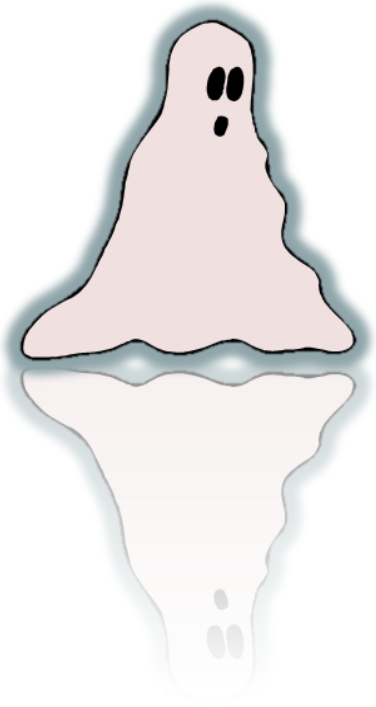
1. Sometimes they do not work and produce more biased results than non-transformed data
2. In most cases, they make interpretation of results more difficult
3. Difficulties to communicate findings to gatekeepers and policy makes
4. Difficulties in comparing findings across studies
5. Do not capitalize on the 'natural' distribution of the data resulting in loss of power
6. Sometimes a transformation can address only one of the problems … and make the other problems worse …



EXAMPLE

The between-gender difference in the Box-Cox transform of weekly minutes of walking for transportation was not significant. **A unit increase on the scale of land-use mix was associated with a 1.82 (95% CI: 0.83, 2.81) increase in the Box-Cox transform of weekly minutes of walking for transportation**.

$$BC\ transform = \frac{(\frac{min}{wk} + 1)^{0.35-1}}{0.35}$$

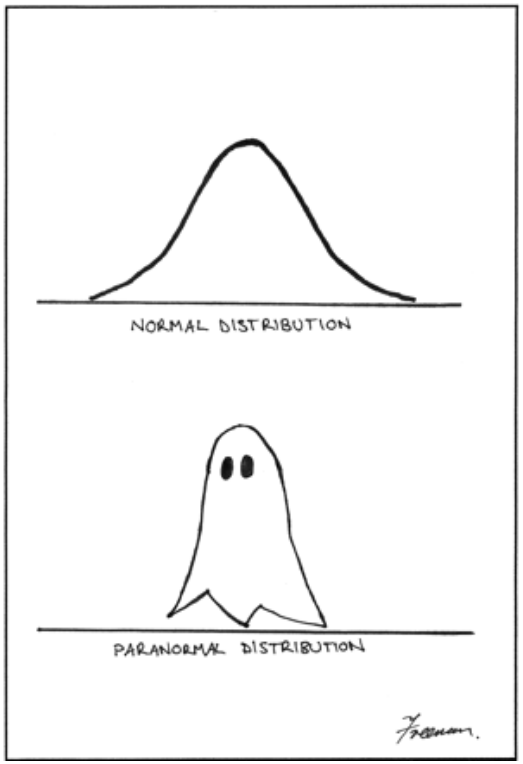


ANOTHER SOLUTION!!

GENERALIZED … MODELS

WHAT ARE THEY?

THE ‘GENERALIZED’ IN GENERALIZED ADDITIVE MIXED MODELS: DEALING WITH “NON-NORMAL” RESIDUALS



G...Ms: WHAT ARE THEY? A BIT OF THEORY

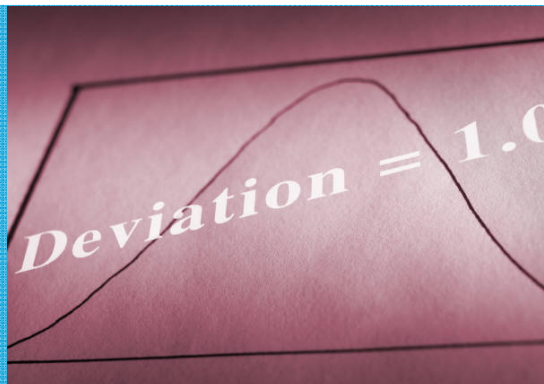
In a G...M, the outcome variable, Y, is assumed to be generated from a particular distribution in the exponential family, a large range of probability distributions that includes the Normal, binomial and Poisson distributions, among others. The mean, μ, of the distribution depends on the independent variables, X, through:

$$E(Y) = \mu = g^{-1}(X\beta)$$

- E(Y) is the expected (mean) value of Y
- Xβ is the linear predictor of unknown parameters (regression coefficients) β
- g is the link function

The above means that:

$$g(\mu) = X\beta$$



G...Ms: WHAT ARE THEY? MORE THEORY ...

G...Ms are made up of 3 components:

1. **Random Component (Variance Function)**

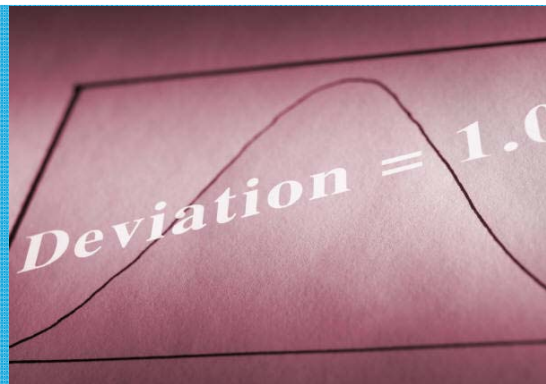
   Identifies dependent variable (Y) and its probability distribution

2. **Systematic Component**

   Identifies the set of explanatory variables ($X_1,...,X_k$)

3. **Link Function**

   Identifies a function of the mean that is a linear function of the explanatory variables

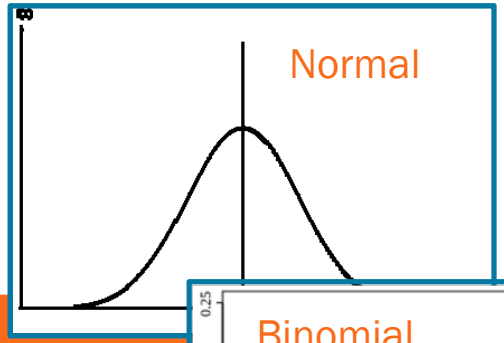


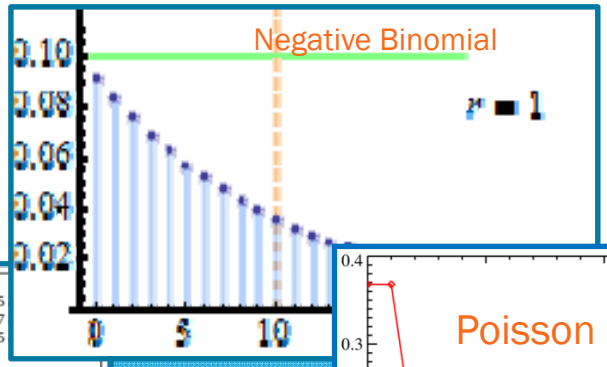
RANDOM COMPONENT or VARIANCE FUNCTION

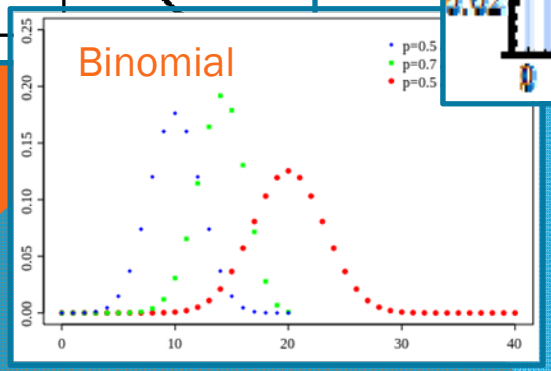
Rather than transforming the data to get approximate normality,

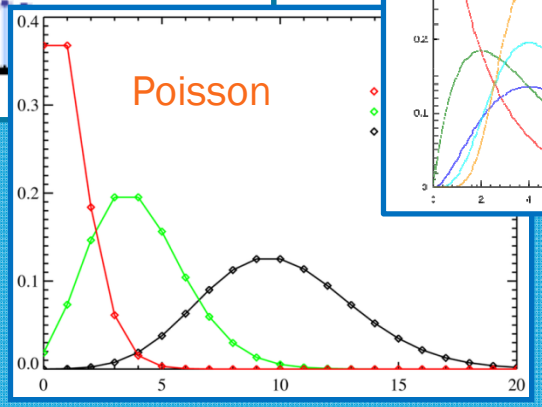
G...Ms expand the allowed distributions for untransformed

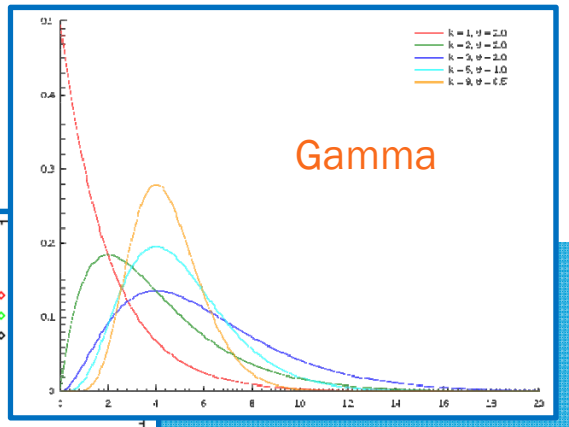
outcomes.











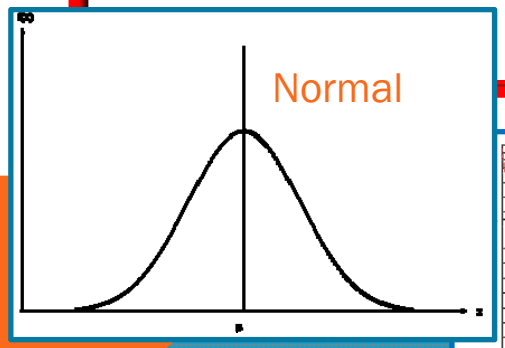
RANDOM COMPONENT or VARIANCE FUNCTION

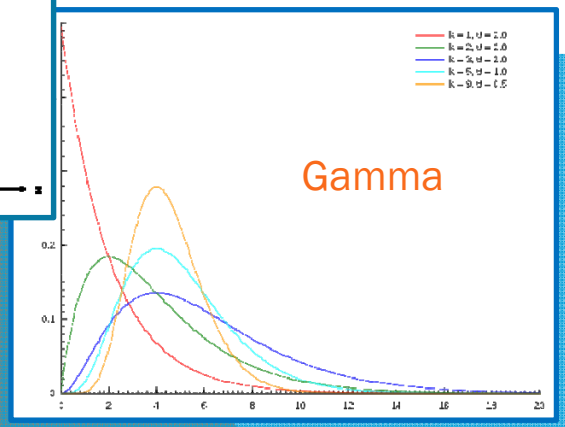
For all these distributions, the variance is the function of the mean (μ):

$$Var(y) = \emptyset Var(\mu)$$

ϕ is called the dispersion parameter and Var(μ) is called the variance function

For the normal distribution $\emptyset = \sigma^2$ and $Var(\mu) = 1$ … the variance **does not** depend on the mean … for all other distributions: **it does.**





HOW TO PICK THE RANDOM COMPONENT??

Normal (Gaussian):

Normally-distributed outcome with constant variance $Var(y) = \sigma^2$

Is a continuous distribution of real numbers ranging from negative to positive infinity

Binomial:

Binary outcomes (e.g., normal weight vs. overweight).

$Var(y) = n\mu(1-\mu)$ $Var(y) \propto n\mu(1-\mu)$ $Var(y) = p(1-p)$

Poisson:

Count data (number of events in fixed area and/or length of time). $Var(\mu) = \mu$ $Var(\mu) \propto \mu$

Negative Binomial:

Count data where $Var(y) > \mu$ $Var(y) = \mu + \mu^2/k$

Gamma:

Continuous positive data with skewed distribution and variation that increases with the mean (variance proportional to the square of the mean; i.e., constant coefficient of variation). The model is robust to wide variation of this assumption. $Var(y) \propto \mu^2$

LINK FUNCTIONS

In classical linear model, the systematic effects of the explanatory variables are assumed to combined additively.

$$E(y) = \mu = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

In G...Ms, this linear combination is also assumed and produces the linear predictor η.

$$\eta = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 = g(\mu)$$

Link function

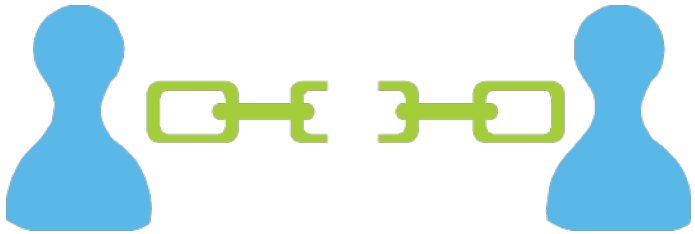
E.g., with log-linear models in which the effects are assumed to be multiplicative on the μ (original) scale, the link function is log : η = log (μ) produces effects that combine additively on the η scale.

Note that the individual data are not transformed to achieve linearity, the means μ are transformed.

The choice of the variance function to model the random component is entirely separate from the choice of link function to achieve linearity of the systematic effects. Thus, a single transformation is not longer trying to do several jobs.

... identity ... logarithmic ... logit ... others ... identity...

LINK FUNCTIONS



Most useful for us:

identity	$\eta = \mu$
log	$\eta = ln(\mu)$
logit	$\eta = ln(\frac{\pi}{1-\pi})$

HOW TO PICK THE LINK FUNCTION … (SOMEWHAT ARBITRARY)

Identity	$\eta = \mu$	Normal; Gamma; sometimes others …
log	$\eta = ln(\mu)$	Normal; Gamma; Poisson; Negative Binomial; Binomial Generally used for outcomes that can take only positive values (discrete or continuous)
logit	$\eta = ln(\frac{\pi}{1-\pi})$	Binomial Used for logistic regression

Decision also based on model fit and analysis of residuals (check linearity).

Interpretation of regression coefficients from various G…Ms

Identity link function and Normal or Gamma distributions:

Amount of change in outcome (in its original units) followed by 1 unit increase in the predictor

Logit link function and Binomial distribution: (antilog of regression coefficients)

Odds ratio – proportional change in odds followed by 1 unit increase in the predictor (>1 = increase or positive association; <1 = decrease or negative association)

Log link function and Normal or Gamma distribution: (antilog of regression coefficients)

Proportional change in outcome followed by 1 unit increase in predictor (>1 = increase or positive association; <1 = decrease or negative association)

Log link function and Poisson or Negative Binomial: (antilog of regression coefficients)

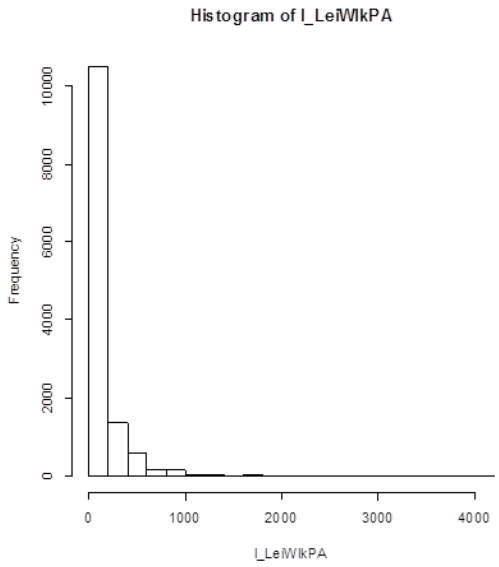
Proportional change in outcome followed by 1 unit increase in predictor (>1 = increase or positive association; <1 = decrease or negative association)

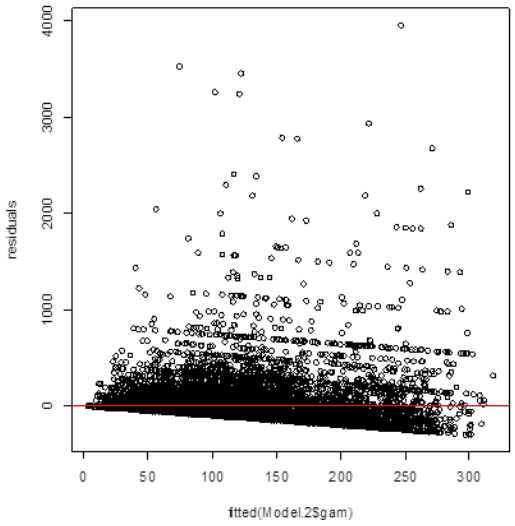
Sometimes called *incidence risk ratio (IRR)* . E.g., IRR=2 means the outcome is twice as prevalent if the associated predictor is increased by one.

EXAMPLE … LET'S GO BACK THE DATA ON WEEKLY MINUTES OF WALKING FOR RECREATION

STEP 1: CHOOSING A VARIANCE FUNCTION

- Data can only take positive values (0 or more)
- Recorded data are discrete – minutes per week; but , in reality, are continuous (time is a continuous variable)
- Data are positively skewed
- The variance of the outcome increases with its mean, the variance is a larger than the mean and squared mean





stats	I_LeiWlkPA
mean	114.8585
sd	224.0988
variance	50220.27
Mean^2	13192.48

EXAMPLE … LET'S GO BACK THE DATA ON WEEKLY MINUTES OF WALKING FOR RECREATION

STEP 1: CHOOSING A VARIANCE FUNCTION

YOUR ANSWER?

Gamma

Negative Binomial

EXAMPLE … LET'S GO BACK THE DATA ON WEEKLY MINUTES OF WALKING FOR RECREATION

STEP 2: CHOOSING A LINK FUNCTION

- Data can only take positive values (0 or more)
- Our variance functions are Gamma or Negative Binomial

Identity?

or

Log?

Log: safer as it cannot produce negative values and stabilizes variance

Identity: easier to interpret and is unlikely to fit the data well

Conclusion: try log

EXAMPLE … GO BACK THE DATA ON WEEKLY MINUTES OF WALKING FOR RECREATION

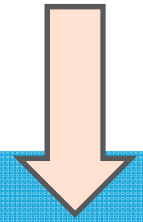
Setting up GAMMs in 'R'

Gamma variance and log link function

```
Gamma.log <- gamm(l_LeiWlkPA+1 ~ fSES + fgender + feducation + fjob + fmarital + Age_final + fcity
    + GN_ResidDen, data=data, family=Gamma(link="log"), random=list(cluster=~1))
```

Negative Binomial variance and log link function

```
NBin.log <- gamm(l_LeiWlkPA+1 ~ fSES + fgender + feducation + fjob + fmarital + Age_final + fcity +
    GN_ResidDen, data=data, family=negative.binomial(1), random=list(cluster=~1))
```



family=

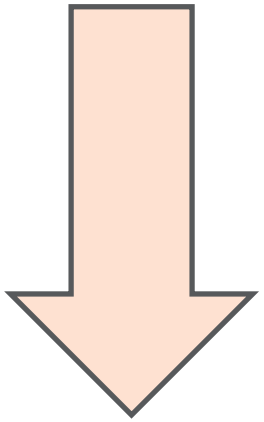
link=

EXAMPLE … GO BACK THE DATA ON WEEKLY MINUTES OF WALKING FOR RECREATION

Setting up G…Ms in R … few more details

Family name (default link function)

```
binomial(link = "logit")
gaussian(link = "identity")
Gamma(link = "inverse")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
```



family =

link =

MODEL SELECTION … WHICH MODEL IS BEST?

Comparing models with different link functions but equal random and systematic components

Consider the AIC or BIC: the model with the smaller values for these parameters are the preferred models

AIC	✓	**Akaike Information Criterion** measure of model fit; index of information lost associated with a model; includes a penalty for increase in parameters (e.g., predictors)
BIC		**Bayesian Information Criterion** As AIC but with greater penalty for increase in parameters

Comparing models with different variance functions and same systematic and link components

Consider the AIC or BIC: the model with the smaller values for these parameters are the preferred models.

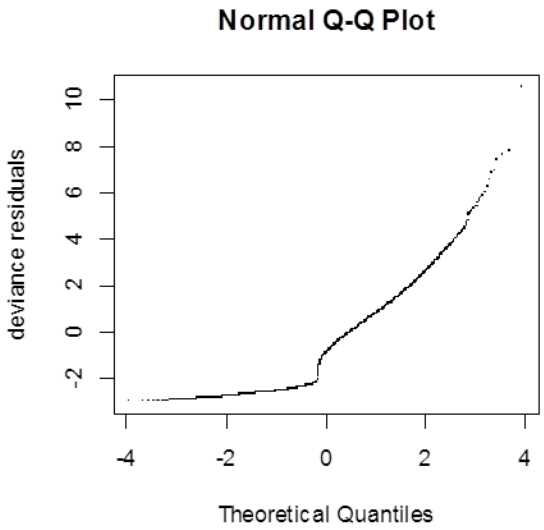
MODEL 2 … WALKING FOR RECREATION (cities as fixed effects)

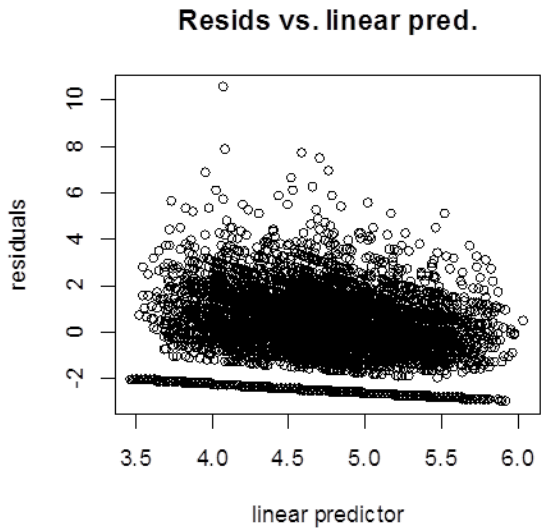
```
GAMMA Variance function
Linear mixed-effects model fit by maximum likelihood
 Data: data
      AIC      BIC    logLik
  51739.5 51940.97 -25842.75
...
Fixed effects: list(fixed)
                                    Value  Std.Error    DF  t-value p-value
XGN_ResidDen                     0.001087 0.00022813 12475  4.76277  0.0000


NEGATIVE BINOMIAL Variance function
Linear mixed-effects model fit by maximum likelihood
 Data: data
       AIC      BIC    logLik
  51737.25 51938.71 -25841.63
…
Fixed effects: list(fixed)
                                    Value  Std.Error    DF  t-value p-value
XGN_ResidDen                     0.001086 0.00022791 12475  4.76547  0.0000
```

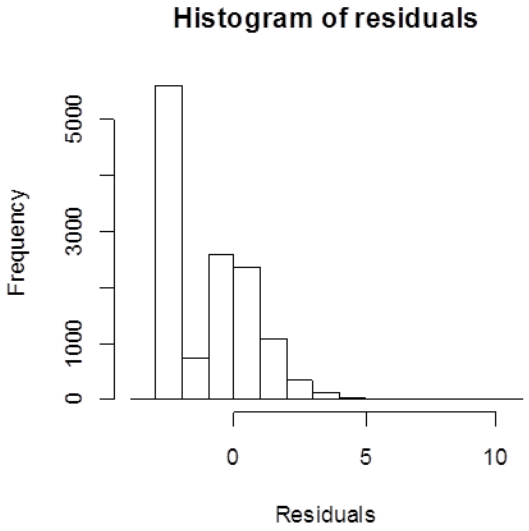
TASK 1:
Set up and run a script with Poisson variance function? Is the model better than the above?

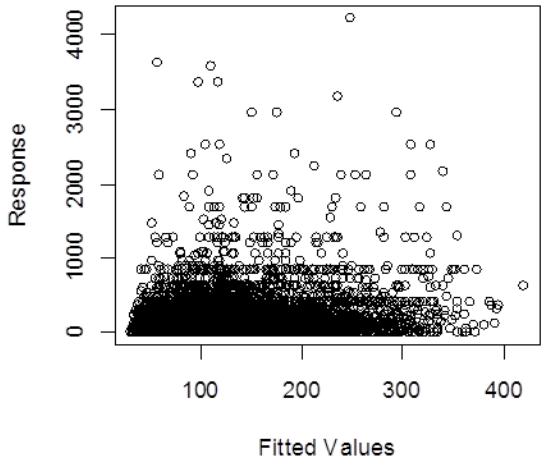
TASK 2:
Calculate the antilog of the regression coefficient and 95% CIs. Explain what you've found.

DIAGNOSTICS … LOOKING AT PLOTS … 'R'









Negative Binomial with log link

A STEP FURTHER … BUMPING INTO NAUGHTY NAUGHTS …

A STEP FURTHER … DEALING WITH NAUGHTY NAUGHTS …

Two-part models

They deal with outcome variables having more zeros than allowed by the distributional assumptions of Gaussian, Gamma, Poisson or Negative Binomial GAMMs.

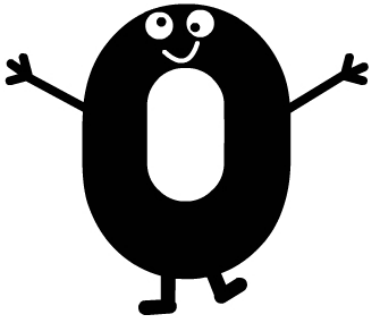
They partition data into two groups:

1. A **binary process** that generates two values (zero vs. non-zero values)

   This can be modelled using a **GAMM with binomial variance function and logit link function** (… as a logistic regression)

2. **Non-zero process** (modelling non-zero values)

   A zero-truncated model based on an appropriate **GAMMs for count or continuous data (Poisson, Negative Binomial, Gaussian or Gamma; with identity or log link functions)**



SETTING UP TWO-PART MODELS IN 'R'

TASKS

1. Create new binary outcome variable

```
complete$LeiWlkDich <- ifelse(I_LeiWlkPA == 0, 0, 1)
```

2. Model it using a GAMM with binomial variance and logit link functions

3. Find an appropriate GAMM model for non-zero values of the outcome variable

OUTPUT OF TWO-PART MODEL IN 'R' (1)

BINARY PROCESS

```
Linear mixed-effects model fit by maximum likelihood
 Data: data
       AIC      BIC    logLik
  55891.08 56092.55 -27918.54

Random effects:
 Formula: ~1 | cluster
        (Intercept)  Residual
StdDev:   0.2716545 0.9932228

Fixed effects: list(fixed)
                          Value  Std.Error    DF   t-value p-value
XGN_ResidDen          0.0008823 0.00027309 12475  3.230821  0.0012
```

```
exp(Dichot.model$lme$coefficients$fixed["XGN_ResidDen"])

XGN_ResidDen
    1.000883
```

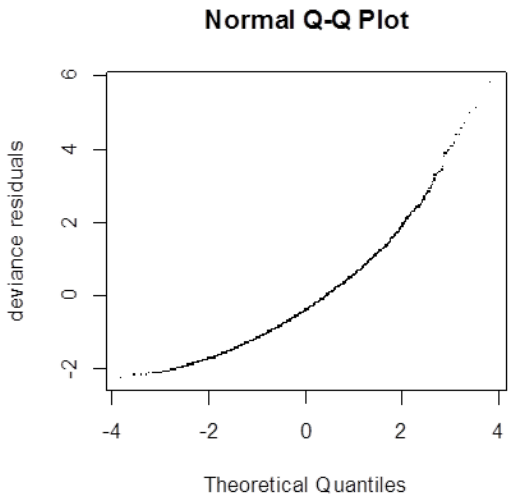
ODDS of Engaging in recreational walking

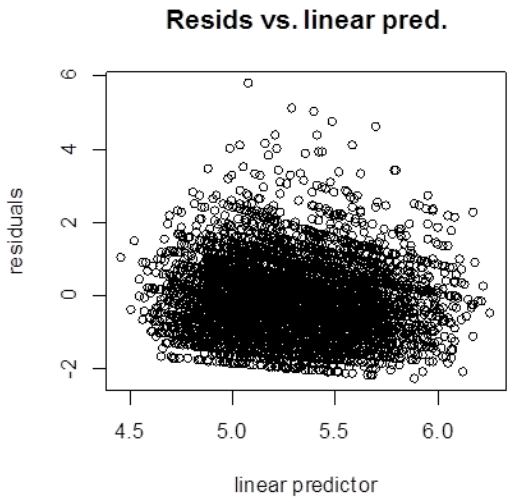
OUTPUT OF TWO-PART MODEL IN 'R' (2)

NON-ZERO PROCESS

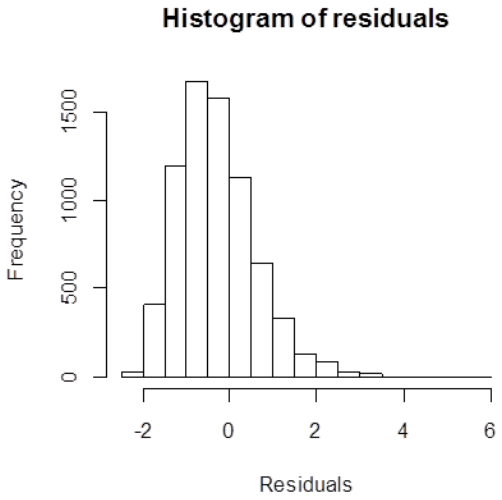
```
Linear mixed-effects model fit by maximum likelihood
 Data: data
       AIC      BIC    logLik
  22958.31 23144.39 -11452.15

Random effects:
 Formula: ~1 | cluster
        (Intercept) Residual
StdDev:   0.0938119 1.161616

Fixed effects: list(fixed)
                              Value  Std.Error   DF  t-value p-value
XGN_ResidDen               0.000731 0.00018454 6899  3.95982  0.0001

-----------------------------------------------------
exp(NZero.model$lme$coefficients$fixed["XGN_ResidDen"])

XGN_ResidDen
    1.000731 How do you interpret this?
```

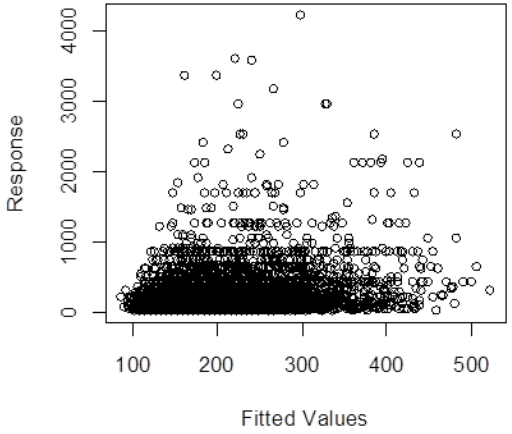
IS THE MODEL VALID? DIAGNOSTICS ...



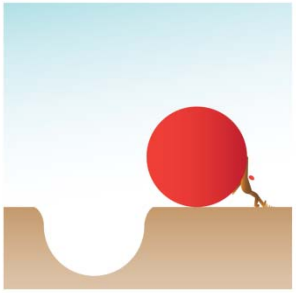


Non-zero process

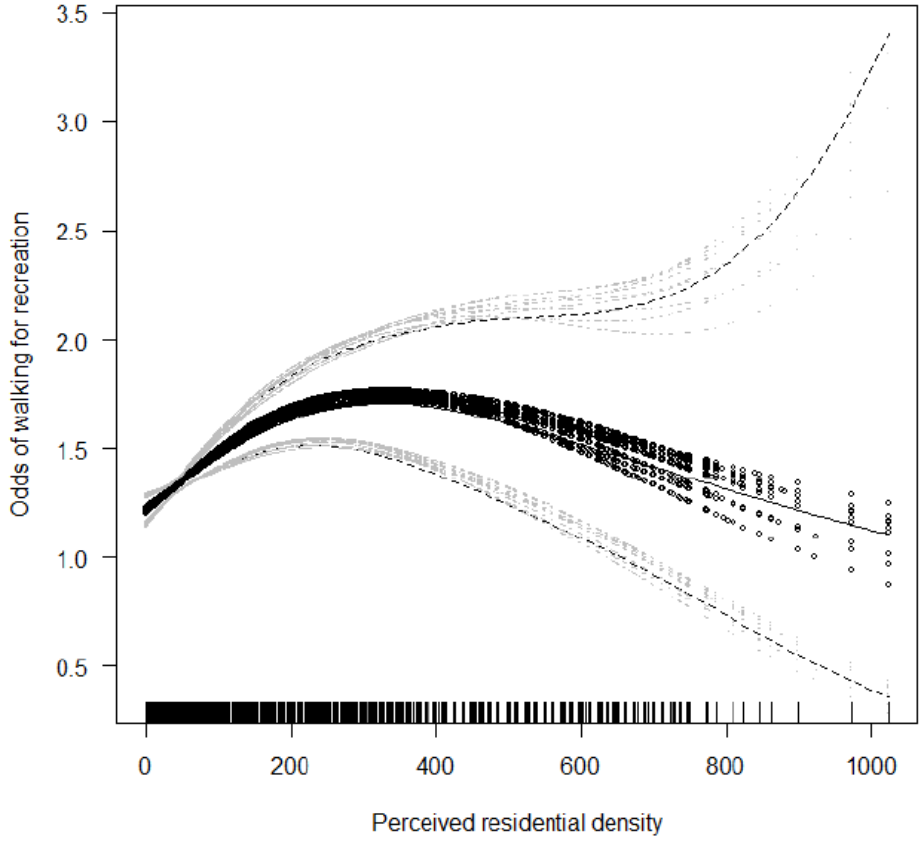




CHALLENGE 3



Non-linear relationships



BAD NEWS: can't simply enter 'untransformed' predictors in the regression models …

IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

THE 'ADDITIVE' IN GENERALIZED ADDITIVE MIXED MODELS: DEALING WITH CURVILINEARITY

How do we estimate this?

- Apply a smoothing model with the odds of walking for recreation as the outcome and Residential Density as a smoother (semi-parametric function)

$$Odds.Walking_{ij} = \beta_{0j} cons + f(R.Density)_{ij} + \beta_2 (SES)_j + \beta_3 (Cities)_j + \beta_4 (\text{cov})_{ij}$$

- Thin plate regression splines
  - Can smooth any number of covariates
  - No knots needed
  - Other optimal statistical properties

IS RESIDENTIAL DENSITY CURVILINEARLY RELATED TO WALKING FOR RECREATION?

IS RESIDENTIAL DENSITY CURVILINEARLY RELATED TO WALKING FOR RECREATION?

Previous model
'Linear' relationship

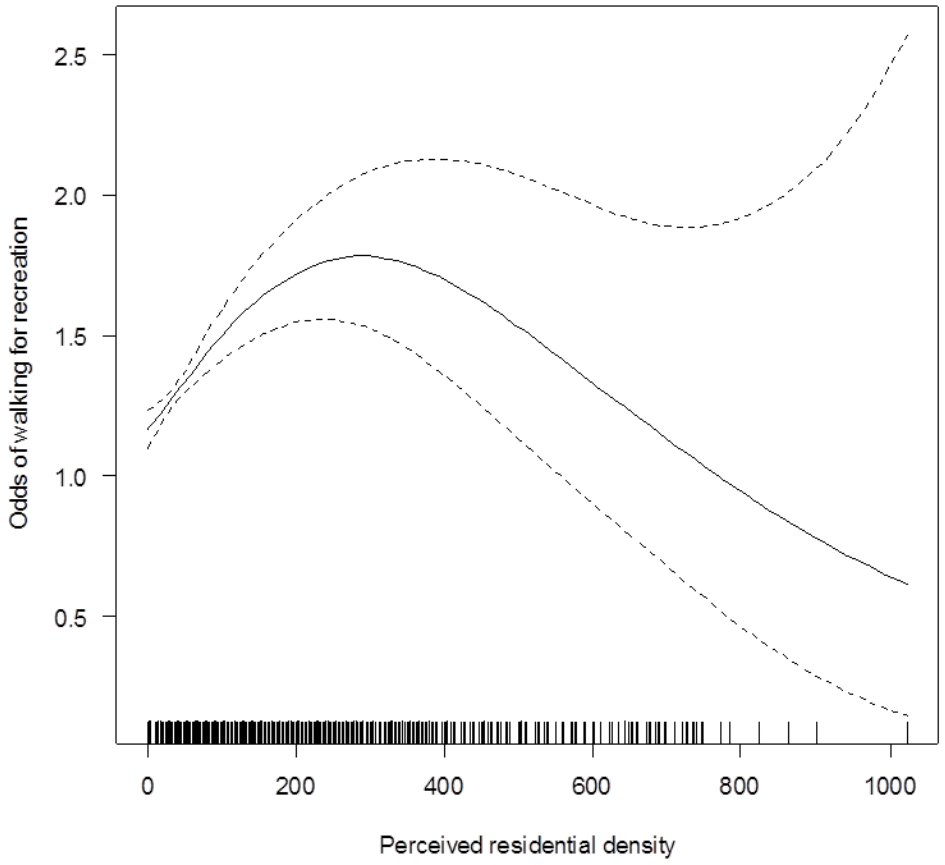
```
Family: binomial
Link function: logit

Formula:
LeiWlkDich ~ fSES + fgender + feducation + fjob + fmarital +
    Age_final + fcity + GN_ResidDen

Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
GN_ResidDen     0.0008823  0.0002731   3.231 0.001237 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0815  Scale est. = 0.98649   n = 12856
```

IS RESIDENTIAL DENSITY CURVILINEARLY RELATED TO WALKING FOR RECREATION?

New model
Curvilinear relationship

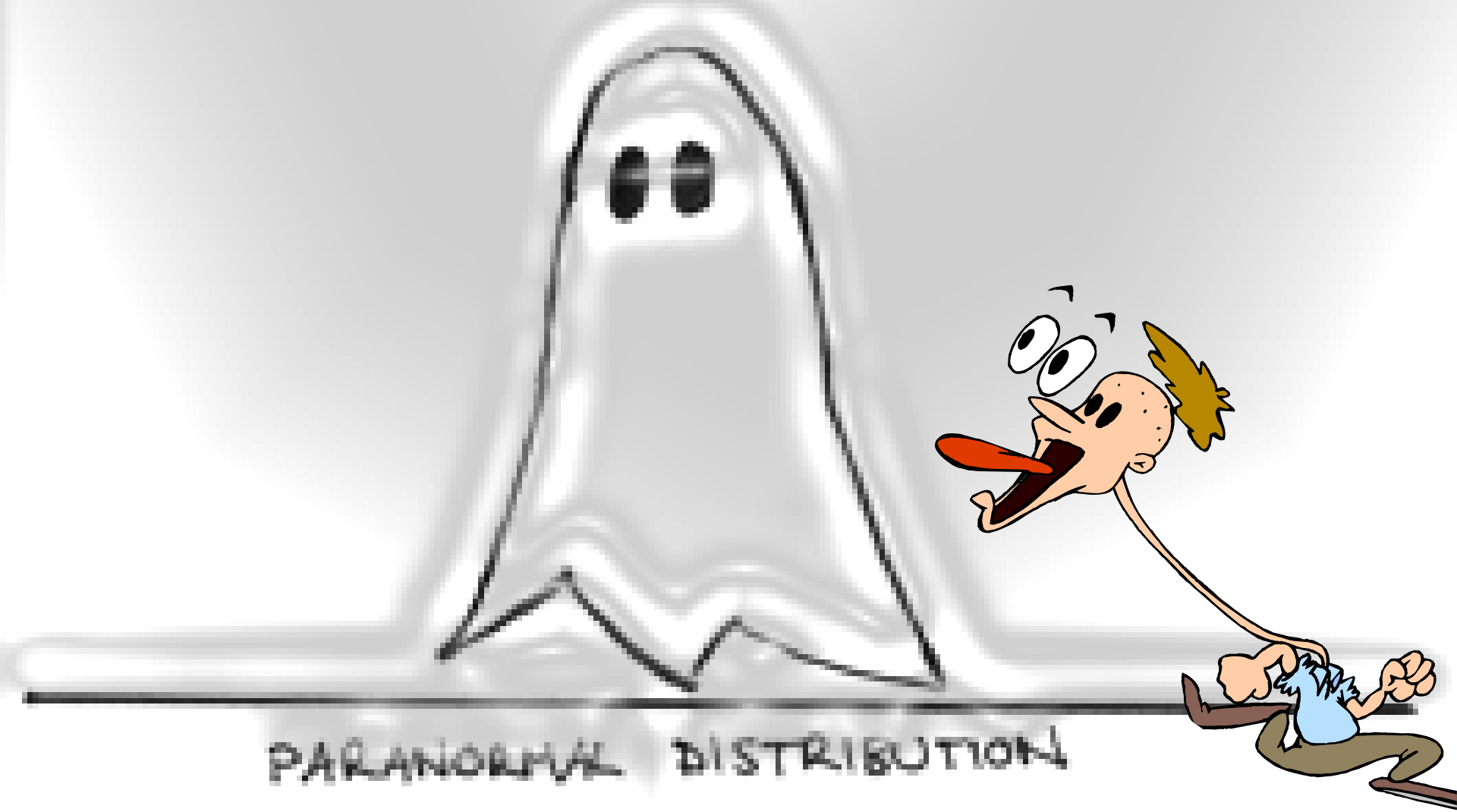
```
Family: binomial
Link function: logit

Formula:
LeiWlkDich ~ fSES + fgender + feducation + fjob + fmarital +
    Age_final + fcity + s(GN_ResidDen)


Approximate significance of smooth terms:
                  edf Ref.df     F  p-value
s(GN_ResidDen) 2.866  2.866 10.48 1.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0836  Scale est. = 0.98725   n = 12856
```

IS RESIDENTIAL DENSITY CURVILINEARLY RELATED TO WALKING FOR RECREATION?



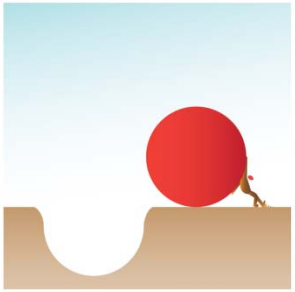

```
plot(Dichot.model2$gam, las=1, ylab="Odds of walking for recreation", se=TRUE,
xlab="Perceived residential density", trans=function(x)exp(x),
shift=mean(predict(Dichot.model2$gam)), page=1)
```

... NOW IT'S YOUR TURN ...



... TEST CURVILINEARITY OF RELATIONSHIP FOR THE NON-ZERO VALUES NEGATIVE BINOMIAL MODEL ...

CHALLENGE 4



Between- and within-city associations

- The strength of these may differ
- Need to assess within- and between-city effects

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 \overline{X}_j$$

Within-city

Between-city

BUT BAD NEWS: we cannot easily distinguish between- and within-city effects …

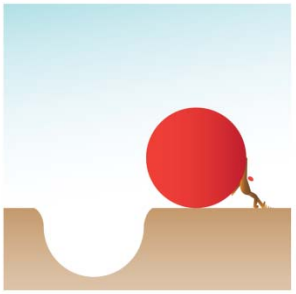
IPEN "sampled" only 17 cities

Cities are not truly "random" factors

- Regression models encompass many predictors

IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK

CHALLENGE 4

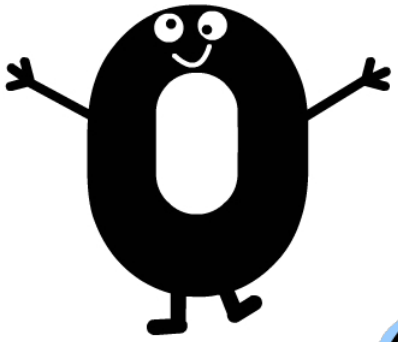


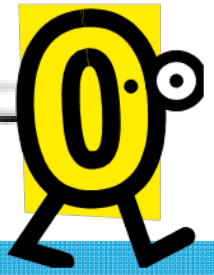
Between- and within-city associations

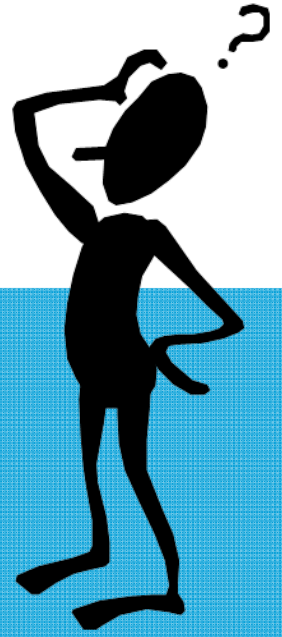
BRAIN STORMING EXERCISE

What can we do about this?

IPEN
INTERNATIONAL PHYSICAL ACTIVITY AND THE ENVIRONMENT NETWORK







THANK YOU!