Linear Regression Module III: Deep Dive

Dr. Mark Williamson

DaCCoTA

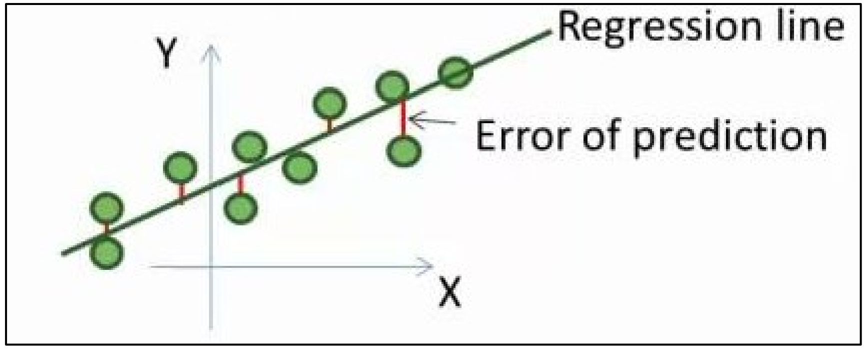
University of North Dakota

Introduction

- Previously:
  - Covered a broad overview
  - Looked at more detail
  - Ran through examples
- This time: looked at more advanced linear regression methods
  - Generalized Linear Mixed Model
  - Longitudinal Analysis
  - Structural Equation Modeling

Reviewing the Basics

- Linear regression: modeling the relationship between a response variable and one or more predictor variables
  - Structure->simple, multiple, multivariate
  - Predictor variables->polynomial, fixed/random, nested
  - Response variables->Gaussian, Logistic, Poisson, etc.
  - Other considerations
- Process of ordinary least squares
- Need to consider assumptions and model fit
- Lots of ways to run a regression



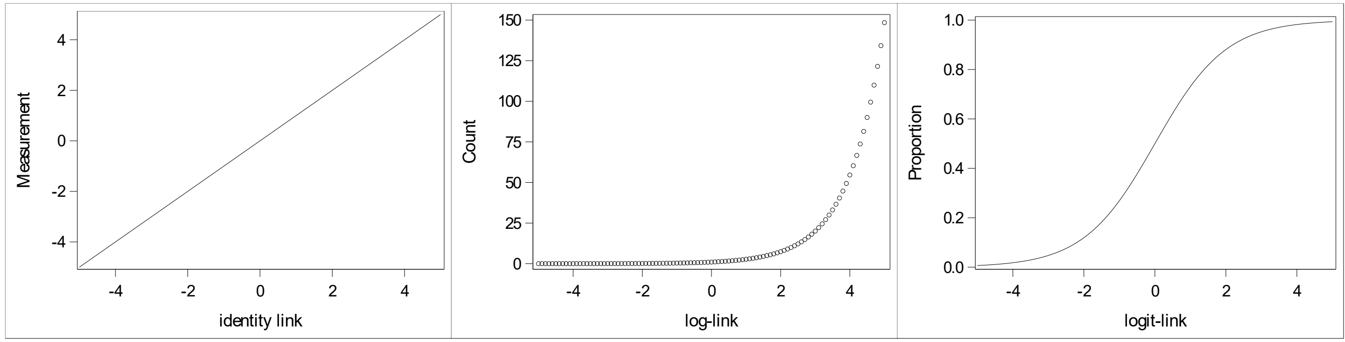
Topics Covered

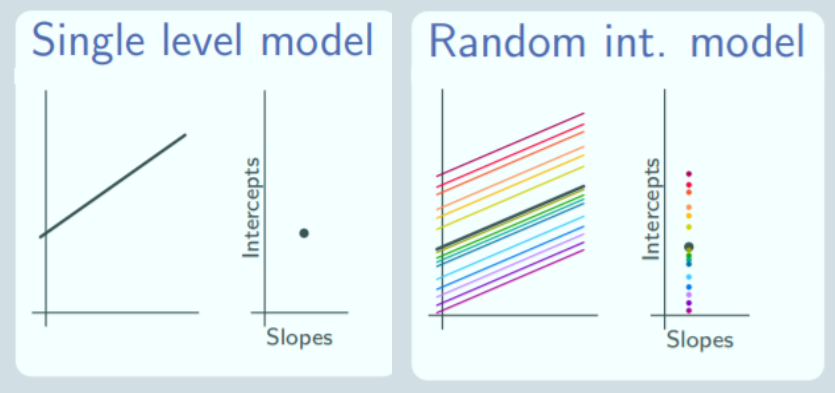
- Generalized Linear Mixed Model
  - Software: SAS Studio
- Longitudinal Analysis
  - Software: R
- Structural Equation Modeling
  - Software: STATA

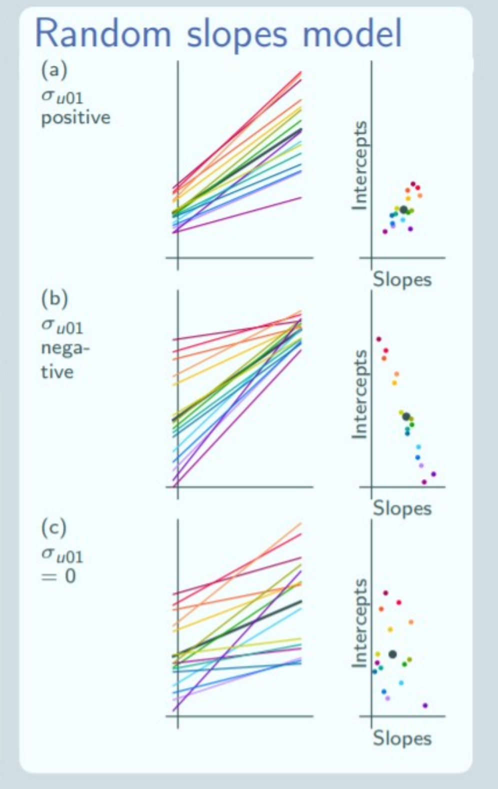
Generalized Linear Mixed Models

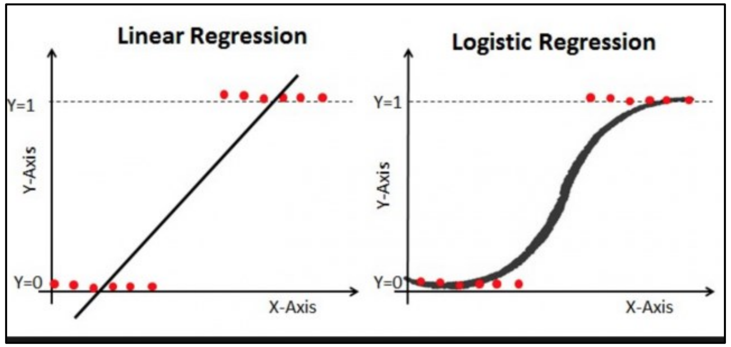
Descriptions

- A Generalized Linear Mixed Model is combination of a Generalized Linear Model and a Linear Mixed Model
  - Generalized -> accommodates non-normal distributions
  - Mixed -> allows for random effects
- Key differences in generalized linear mixed model and linear model:
  - Method of estimation: Ordinary Least Squares vs. Maximum Likelihood (iteratively maximize likelihood of parameters given data)
  - Distributions: Normal distribution vs. Others
  - Model scale: GLMMs link expected values to model scale with link
- Random Intercepts/ Random Slopes:









Generalized Linear Mixed Models

Formats

Basic (Random Intercepts):

```
PROC GLIMMIX data=dataset;
        class TREATMENT RANDOM;
        model RESPONSE= TREATMENT;
        random intercept /subject=RANDOM;
PROC GLIMMIX data=dataset;
        class RANDOM;
        model RESPONSE= TREATMENT;
        random intercept /subject=RANDOM;
PROC GLIMMIX data=dataset;
        class RANDOM;
        model RESPONSE= TREATMENT;
        random RANDOM;
```

Nested:

```
PROC GLIMMIX data=dataset;
        class TEACHINGSTYLE STATE SCHOOL;
        model RESPONSE = TEACHINGSTYLE;
        random STATE SCHOOL(STATE);
```

Random Effects Only;

```
PROC GLIMMIX data=dataset;
        class TREE BRANCH LEAF;
        model RESPONSE = ;
        random TREE BRANCH(TREE) LEAF(BRANCH TREE);
```

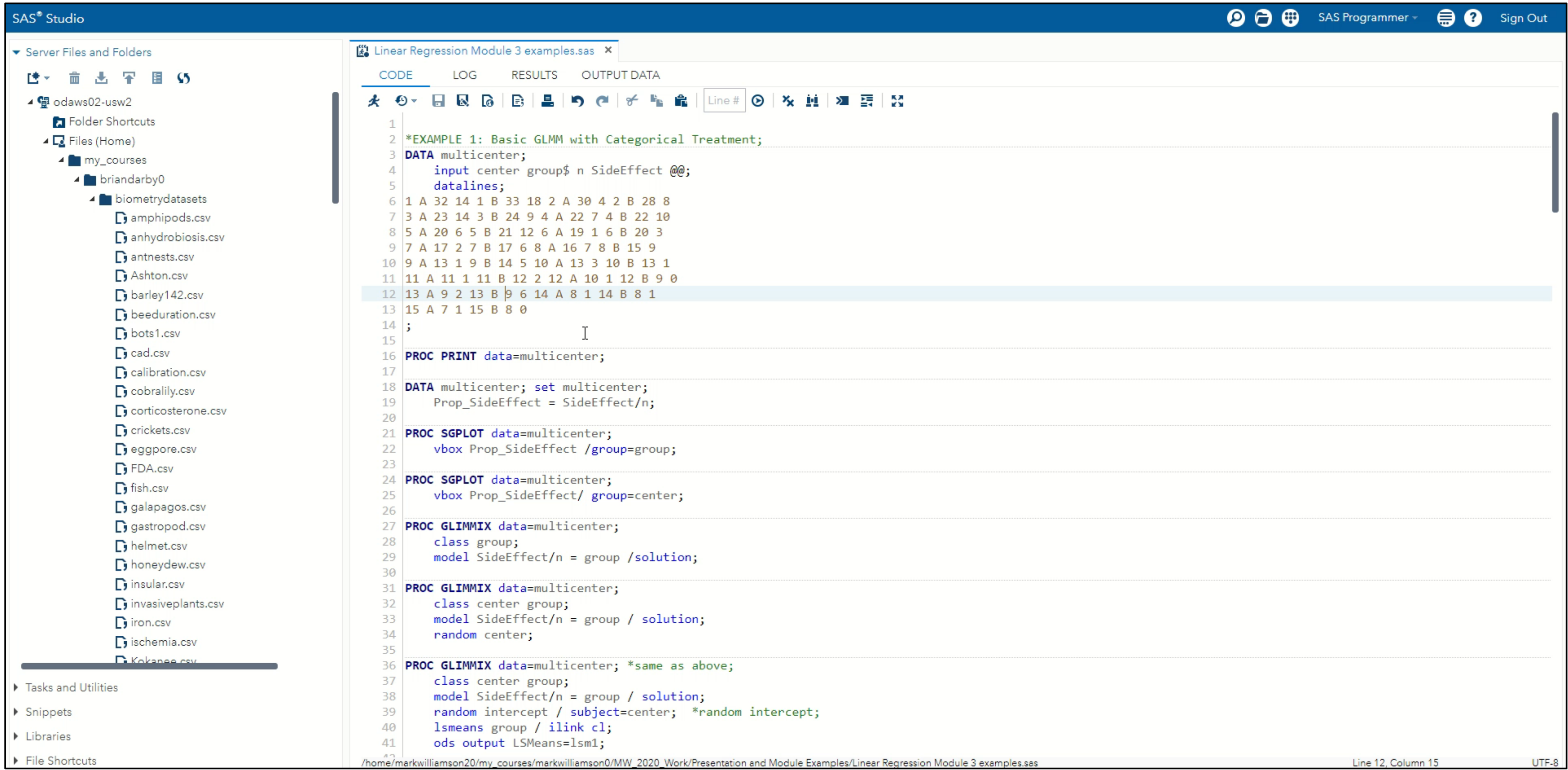
Random Slopes:

```
PROC GLIMMIX data=dataset;
        class RANDOM;
        model RESPONSE = TREATMENT;
        random intercept TREATMENT/subject=RANDOM;
```

Generalized Linear Mixed Models

Examples

- Multicenter[1]
- NPK
- RC[2]
- Wings



```

*EXAMPLE 1: Basic GLMM with Categorical Treatment;
DATA multicenter;
    input center group$ n SideEffect @@;
    datalines;
1 A 32 14 1 B 33 18 2 A 30 4 2 B 28 8
3 A 23 14 3 B 24 9 4 A 22 7 4 B 22 10
5 A 20 6 5 B 21 12 6 A 19 1 6 B 20 3
7 A 17 2 7 B 17 6 8 A 16 7 8 B 15 9
9 A 13 1 9 B 14 5 10 A 13 3 10 B 13 1
11 A 11 1 11 B 12 2 12 A 10 1 12 B 9 0
13 A 9 2 13 B 9 6 14 A 8 1 14 B 8 1
15 A 7 1 15 B 8 0
;

PROC PRINT data=multicenter;

DATA multicenter; set multicenter;
    Prop_SideEffect = SideEffect/n;

PROC SGPLOT data=multicenter;
    vbox Prop_SideEffect /group=group;

PROC SGPLOT data=multicenter;
    vbox Prop_SideEffect/ group=center;

PROC GLIMMIX data=multicenter;
    class group;
    model SideEffect/n = group /solution;

PROC GLIMMIX data=multicenter;
    class center group;
    model SideEffect/n = group / solution;
    random center;

PROC GLIMMIX data=multicenter; *same as above;
    class center group;
    model SideEffect/n = group / solution;
    random intercept / subject=center;  *random intercept;
    lsmeans group / ilink cl;
    ods output LSMeans=lsm1;
```

/home/markwilliamson20/my_courses/markwilliamson0/MW_2020_Work/Presentation and Module Examples/Linear Regression Module 3 examples.sas

[1] https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/196-30.pdf

[2] https://v8doc.sas.com/sashtml/stat/chap41/sect33.htm

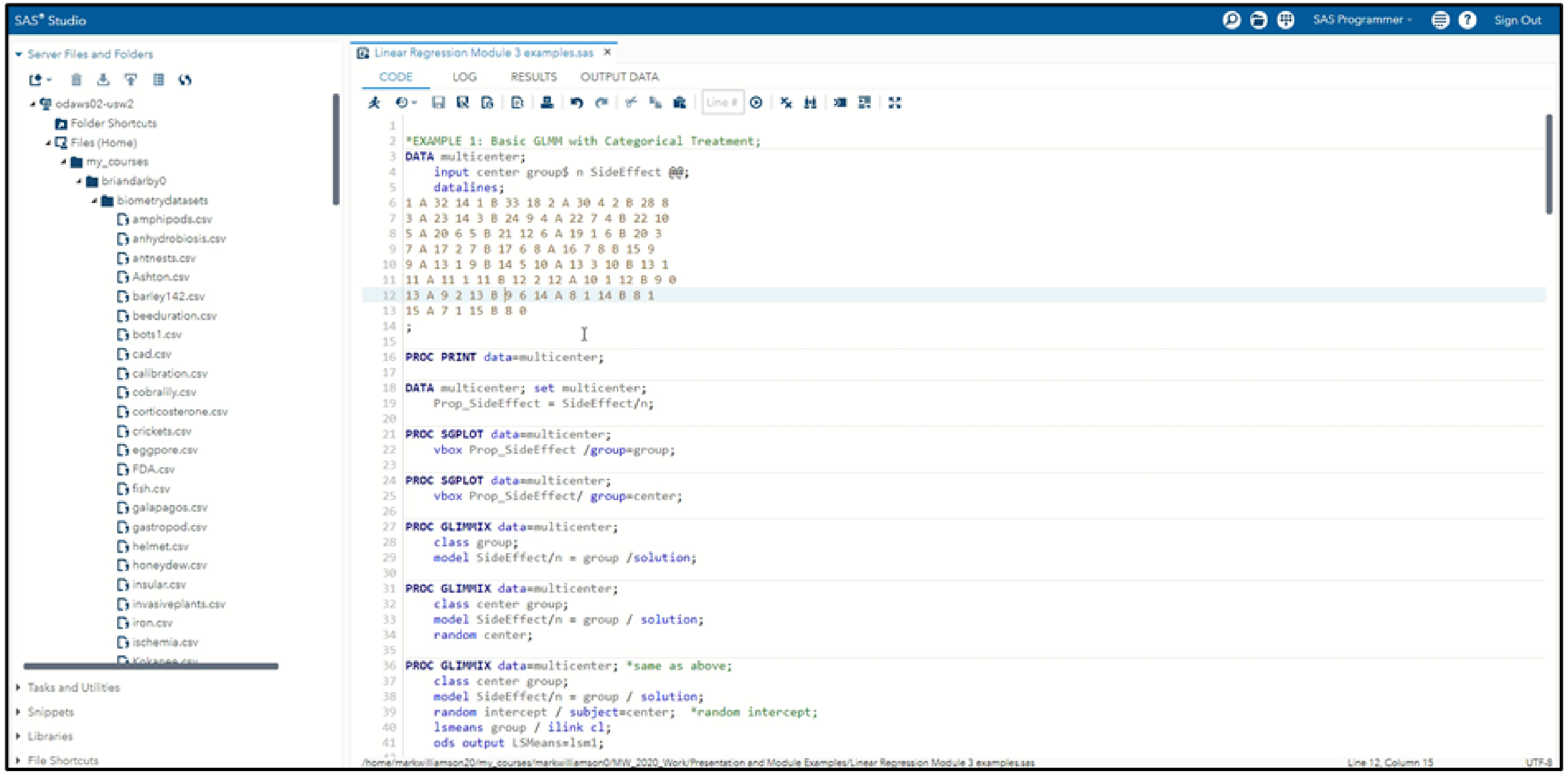
DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Generalized Linear Mixed Models

UND UNIVERSITY OF NORTH DAKOTA

Examples

- Multicenter[1]
- NPK
- RC[2]
- Wings



```
*EXAMPLE 1: Basic GLMM with Categorical Treatment;
DATA multicenter;
    input center group$ n SideEffect @@;
    datalines;
1 A 32 14 1 B 33 18 2 A 30 4 2 B 28 8
3 A 23 14 3 B 24 9 4 A 22 7 4 B 22 10
5 A 20 6 5 B 21 12 6 A 19 1 6 B 20 3
7 A 17 2 7 B 17 6 8 A 16 7 8 B 15 9
9 A 13 1 9 B 14 5 10 A 13 3 10 B 13 1
11 A 11 1 11 B 12 2 12 A 10 1 12 B 9 0
13 A 9 2 13 B 9 6 14 A 8 1 14 B 8 1
15 A 7 1 15 B 8 0
;

PROC PRINT data=multicenter;

DATA multicenter; set multicenter;
    Prop_SideEffect = SideEffect/n;

PROC SGPLOT data=multicenter;
    vbox Prop_SideEffect /group=group;

PROC SGPLOT data=multicenter;
    vbox Prop_SideEffect/ group=center;

PROC GLIMMIX data=multicenter;
    class group;
    model SideEffect/n = group /solution;

PROC GLIMMIX data=multicenter;
    class center group;
    model SideEffect/n = group / solution;
    random center;

PROC GLIMMIX data=multicenter; *same as above;
    class center group;
    model SideEffect/n = group / solution;
    random intercept / subject=center;  *random intercept;
    lsmeans group / ilink cl;
    ods output LSMeans=lsm1;
```

[1] https://support.sas.com/resources/papers/proceedings/proceedings/sugi30/196-30.pdf

[2] https://v8doc.sas.com/sashtml/stat/chap41/sect33.htm

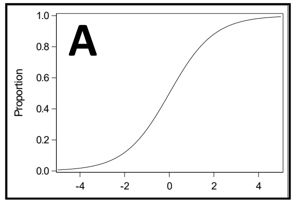
Assessment 1

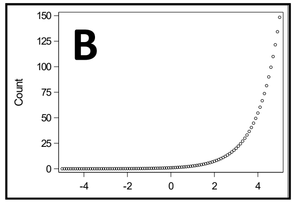
1. You want to run PROC GLIMMIX to model a medical outcome (Med) as a function of a categorical treatment (Treat) with the hospital (Hosp) as a random effect. How would you set up the following SAS code?

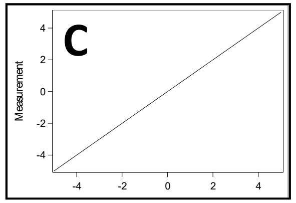
```
PROC GLIMMIX data=dataset;
CLASS ____________;
MODEL ____________ = ____________;
RANDOM intercept /subject=____________;
```

2. Match the following distributions to their appropriate link function (A, B, or C).

1) Poisson 2) Negative Binomial 3) Binary 4) Normal



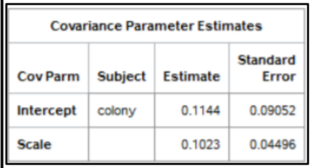




3. Which of the following are likely random effects?

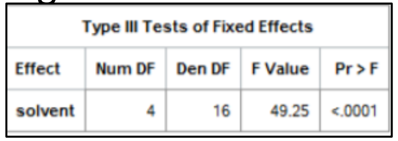
Treatment, School, Ethnicity, Temperature, Block, Color, Chemical concentration, Site

4. Below are tables from a GLIMMIX procedure. Is the fixed effect significant, did the random effect affect the results, and is the model a good fit?



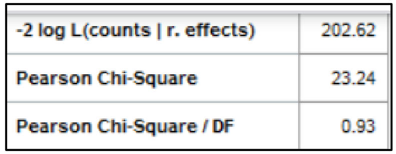
Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	colony	0.1144	0.09052
Scale		0.1023	0.04496



Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
solvent	4	16	49.25	<.0001



-2 log L(counts | r. effects)	202.62
Pearson Chi-Square	23.24
Pearson Chi-Square / DF	0.93

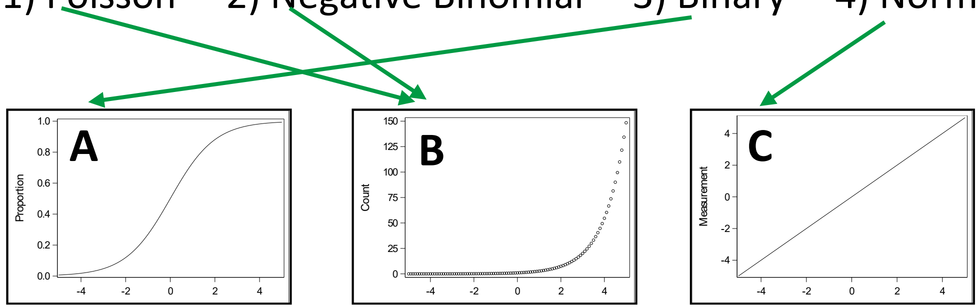
Assessment 1

1. You want to run PROC GLIMMIX to model a medical outcome (Med) as a function of a categorical treatment (Treat) with the hospital (Hosp) as a random effect. How would you set up the following SAS code?

PROC GLIMMIX data=dataset;
CLASS **Treat Hosp**;
MODEL **Med** = **Treat**;
RANDOM intercept /subject=**Hosp**;

2. Match the following distributions to their appropriate link function (A, B, or C).

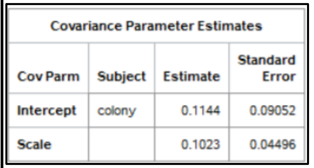
1) Poisson 2) Negative Binomial 3) Binary 4) Normal



3. Which of the following are likely random effects?

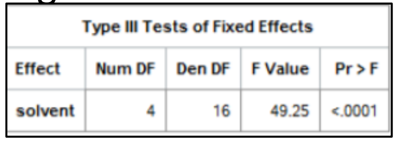
Treatment, **School**, Ethnicity, Temperature, **Block**, Color, Chemical concentration, **Site**

4. Below are tables from a GLIMMIX procedure. Is the fixed effect significant, did the random effect affect the results, and is the model a good fit?



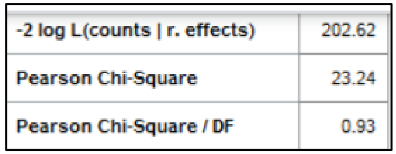
Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error
Intercept	colony	0.1144	0.09052
Scale		0.1023	0.04496



Type III Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
solvent	4	16	49.25	<.0001



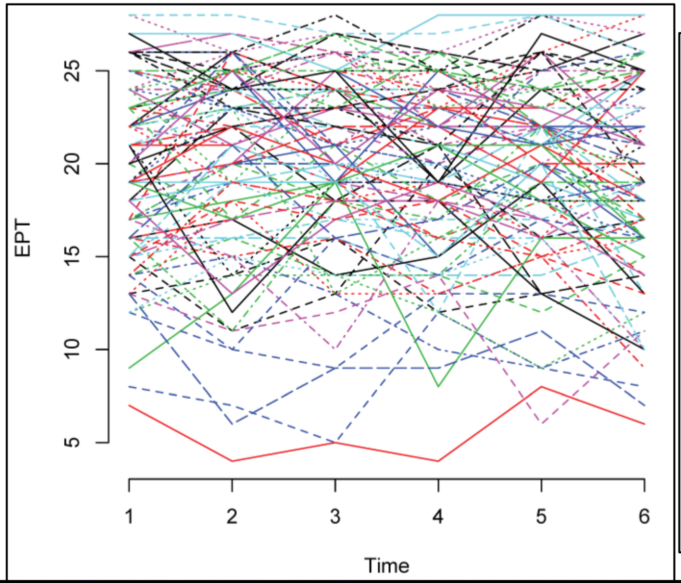
-2 log L(counts | r. effects)	202.62
Pearson Chi-Square	23.24
Pearson Chi-Square / DF	0.93

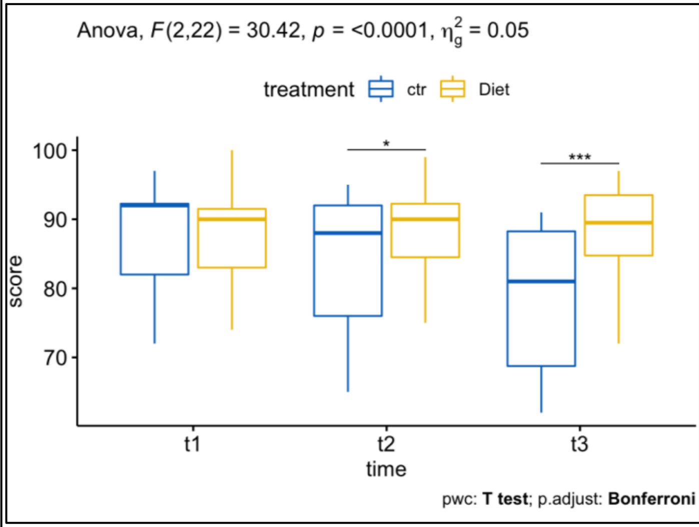
Yes, solvent is significant ($p<0.05$), and colony is non-zero, so it had some effect on the results. The Pearson Chi-Square/DF is around 1.0, so the model is not over dispersed.

Longitudinal Analysis

Descriptions

- Longitudinal data can be viewed as a special case of the multilevel data
- Time is nested within individual participants/observations
- Response variable and predictor variable(s) measured several times
- Point is to characterize change
- Parameters needed to link predictors to response and account for correlational structure of repeated measurements
- Simplest Case: Repeated Measures ANOVA
- Other: Linear mixed effects models, Generalized estimating equations





Longitudinal Analysis

Formats

Repeated Measures ANOVA:

```
aov(Y~TREATMENT*TIME +
        Error(RANDOM),data=DATASET)
```

```
NESTEDDATASET <- groupedData(Y ~ TREATMENT |
        RANDOM, data=DATASET)
```

```
gls(Y ~ TREATMENT*TIME, data=NESTEDDATASET,
    ❖ corr=corCompSymm(, form= ~ 1 | RANDOM))
    ❖ corr=corSymm(, form= ~ 1 | RANDOM),
        weights = varIdent(form = ~ 1 | TIME))
    ❖ corr=corAR1(, form= ~ 1 | RANDOM))
    ❖ corr=corAR1(, form= ~ 1 | RANDOM),
        weights=varIdent(form = ~ 1 | TIME))
```

Linear Mixed Effects:

```
#Random intercept
lmer(Y ~ TREAT + TIME + CAT + (1 | RANDOM), data = DATASET)

#Random intercept and slope
lmer(Y ~ TREAT + TIME + CAT + (TIME | RANDOM), data = DATASET)
```

Generalized Estimating Equations:

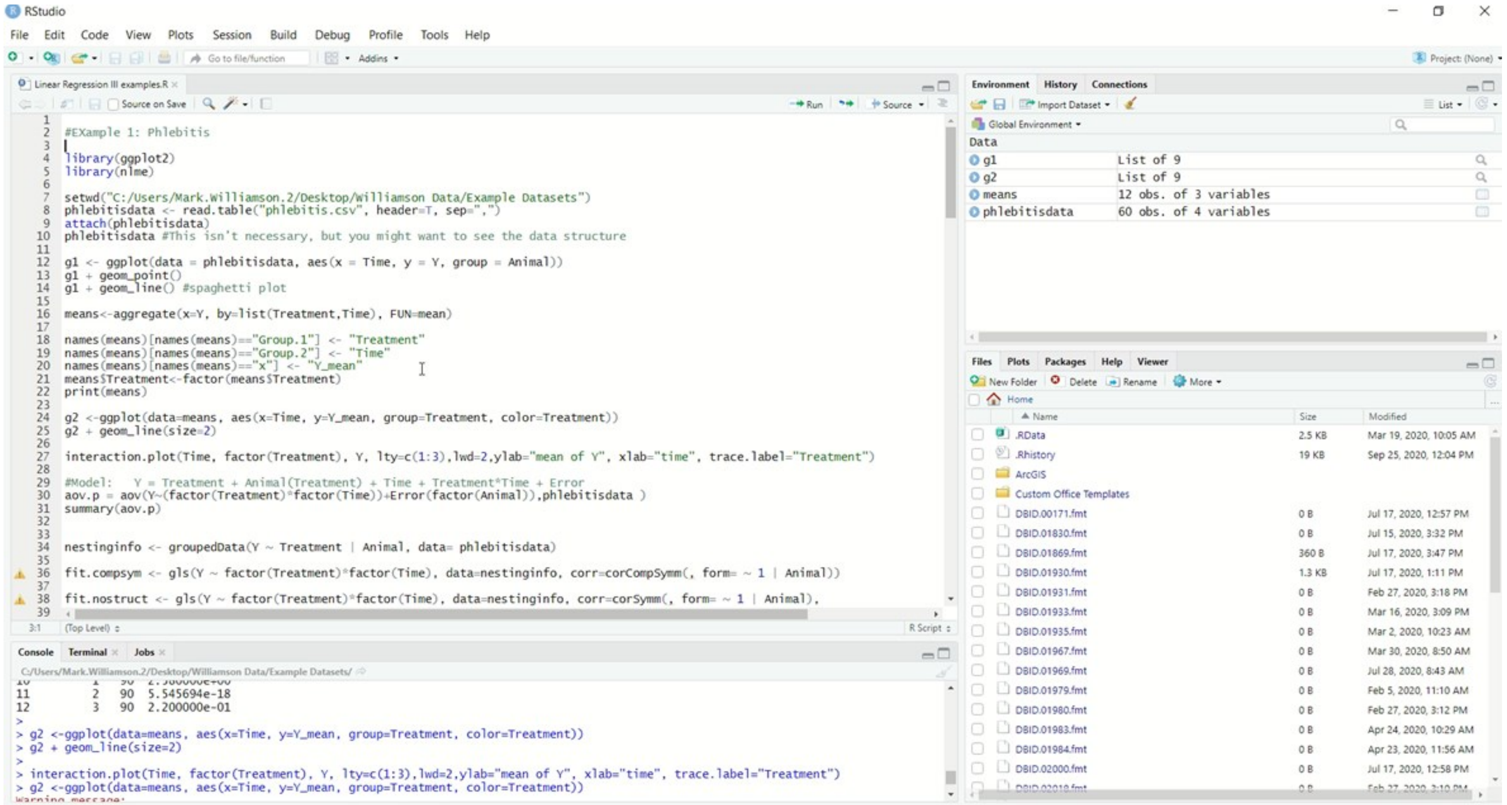
```
glm(Y~ TREATMENT, data=DATASET, family="DISTRUBITION"

gee(Y~ TREATMENT, data=DATASET, family="DISTRIBUTION", id=RANDOM,
    ❖ corstr = "independence", scale.fix = TRUE, scale.value = 1)
    ❖ corstr = "exchangeable",  scale.fix = TRUE, scale.value = 1)
    ❖ corstr = "exchangeable",  scale.fix = FALSE, scale.value = 1)
```

Longitudinal Analysis

Examples

- Phlebitis (RM-ANOVA)[1]
- Beat the Blues (LME)[2]
- Respiratory (GEE)[2]
- Epilepsy (GEE)[2]


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Linear Regression III examples.R
Source on Save
Run
Source
Environment History Connections
Import Dataset
List
Global Environment
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home
(Top Level)
R Script
Console Terminal Jobs

```r
#EXample 1: Phlebitis

library(ggplot2)
library(nlme)

setwd("C:/Users/Mark.Williamson.2/Desktop/Williamson Data/Example Datasets")
phlebitisdata <- read.table("phlebitis.csv", header=T, sep=",")
attach(phlebitisdata)
phlebitisdata #This isn't necessary, but you might want to see the data structure

g1 <- ggplot(data = phlebitisdata, aes(x = Time, y = Y, group = Animal))
g1 + geom_point()
g1 + geom_line() #spaghetti plot

means<-aggregate(x=Y, by=list(Treatment,Time), FUN=mean)

names(means)[names(means)=="Group.1"] <- "Treatment"
names(means)[names(means)=="Group.2"] <- "Time"
names(means)[names(means)=="x"] <- "Y_mean"
means$Treatment<-factor(means$Treatment)
print(means)

g2 <-ggplot(data=means, aes(x=Time, y=Y_mean, group=Treatment, color=Treatment))
g2 + geom_line(size=2)

interaction.plot(Time, factor(Treatment), Y, lty=c(1:3),lwd=2,ylab="mean of Y", xlab="time", trace.label="Treatment")

#Model:   Y = Treatment + Animal(Treatment) + Time + Treatment*Time + Error
aov.p = aov(Y~(factor(Treatment)*factor(Time))+Error(factor(Animal)),phlebitisdata )
summary(aov.p)


nestinginfo <- groupedData(Y ~ Treatment | Animal, data= phlebitisdata)

fit.compsym <- gls(Y ~ factor(Treatment)*factor(Time), data=nestinginfo, corr=corCompSymm(, form= ~ 1 | Animal))

fit.nostruct <- gls(Y ~ factor(Treatment)*factor(Time), data=nestinginfo, corr=corSymm(, form= ~ 1 | Animal),
```

```
Data
g1               List of 9
g2               List of 9
means            12 obs. of 3 variables
phlebitisdata    60 obs. of 4 variables
```

```
C:/Users/Mark.Williamson.2/Desktop/Williamson Data/Example Datasets/
11          2   90  5.545694e-18
12          3   90  2.200000e-01
>
> g2 <-ggplot(data=means, aes(x=Time, y=Y_mean, group=Treatment, color=Treatment))
> g2 + geom_line(size=2)
>
> interaction.plot(Time, factor(Treatment), Y, lty=c(1:3),lwd=2,ylab="mean of Y", xlab="time", trace.label="Treatment")
> g2 <-ggplot(data=means, aes(x=Time, y=Y_mean, group=Treatment, color=Treatment))
```

Name	Size	Modified
.RData	2.5 KB	Mar 19, 2020, 10:05 AM
.Rhistory	19 KB	Sep 25, 2020, 12:04 PM
ArcGIS		
Custom Office Templates		
DBID.00171.fmt	0 B	Jul 17, 2020, 12:57 PM
DBID.01830.fmt	0 B	Jul 15, 2020, 3:32 PM
DBID.01869.fmt	360 B	Jul 17, 2020, 3:47 PM
DBID.01930.fmt	1.3 KB	Jul 17, 2020, 1:11 PM
DBID.01931.fmt	0 B	Feb 27, 2020, 3:18 PM
DBID.01933.fmt	0 B	Mar 16, 2020, 3:09 PM
DBID.01935.fmt	0 B	Mar 2, 2020, 10:23 AM
DBID.01967.fmt	0 B	Mar 30, 2020, 8:50 AM
DBID.01969.fmt	0 B	Jul 28, 2020, 8:43 AM
DBID.01979.fmt	0 B	Feb 5, 2020, 11:10 AM
DBID.01980.fmt	0 B	Feb 27, 2020, 3:12 PM
DBID.01983.fmt	0 B	Apr 24, 2020, 10:29 AM
DBID.01984.fmt	0 B	Apr 23, 2020, 11:56 AM
DBID.02000.fmt	0 B	Jul 17, 2020, 12:58 PM

[1] https://online.stat.psu.edu/stat510/lesson/10/10.1

[2] A Handbook of Statistical Analyses Using R

Assessment 2

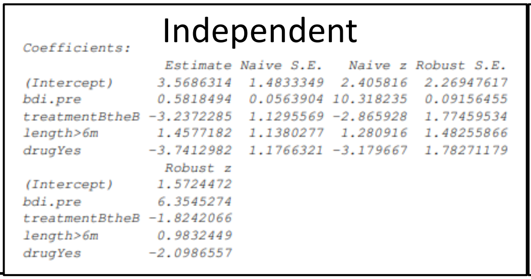
1. In R, which of the codes below correctly codes for a random intercept?

a) (RANDOM) b) (1|RANDOM)

c) (RANOM|1) c) (TIME|RANDOM)

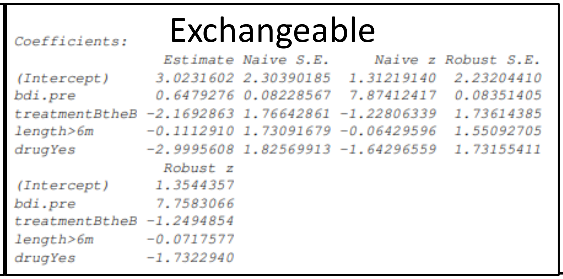
2. What approach would be best to used for Poisson distributed data, a Linear Mixed Effects model or a Generalized Estimating Equation model? Why?

3. Below are summary results for two GEEs with different correlation structures. Which model is a more realistic fit to the data? Why?



Independent

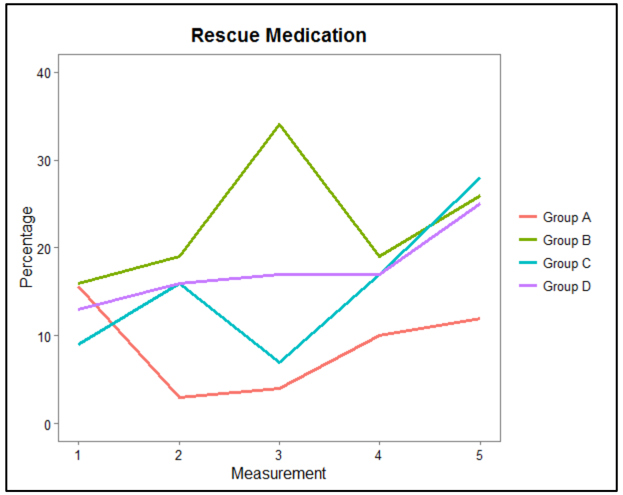
```
Coefficients:
              Estimate Naive S.E.   Naive z Robust S.E.
(Intercept)   3.5686314  1.4833349  2.405816  2.26947617
bdi.pre       0.5818494  0.0563904 10.318235  0.09156455
treatmentBtheB -3.2372285 1.1295569 -2.865928  1.77459534
length>6m     1.4577182  1.1380277  1.280916  1.48255866
drugYes      -3.7412982  1.1766321 -3.179667  1.78271179
              Robust z
(Intercept)   1.5724472
bdi.pre       6.3545274
treatmentBtheB -1.8242066
length>6m     0.9832449
drugYes      -2.0986557
```



Exchangeable

```
Coefficients:
              Estimate Naive S.E.      Naive z Robust S.E.
(Intercept)   3.0231602 2.30390185  1.31219140  2.23204410
bdi.pre       0.6479276 0.08228567  7.87412417  0.08351405
treatmentBtheB -2.1692863 1.76642861 -1.22806339  1.73614385
length>6m    -0.1112910 1.73091679 -0.06429596  1.55092705
drugYes      -2.9995608 1.82569913 -1.64296559  1.73155411
              Robust z
(Intercept)   1.3544357
bdi.pre       7.7583066
treatmentBtheB -1.2494854
length>6m    -0.0717577
drugYes      -1.7322940
```

4. To the right is a spaghetti plot of the percent of patients who took rescue medication by group. Does there appear to be differences across group and time (measurement)?



DaCCoTA
DAKOTA CANCER COLLABORATIVE ON TRANSLATIONAL ACTIVITY

UND UNIVERSITY OF NORTH DAKOTA

Assessment 2

1. In R, which of the codes below correctly codes for a random intercept?

a) (RANDOM) **b) (1|RANDOM)**

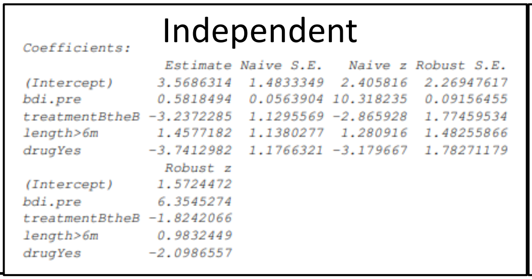
c) (RANOM|1) c) (TIME|RANDOM)

2. What approach would be best to used for Poisson distributed data, a Linear Mixed Effects model or a Generalized Estimating Equation model? Why?

A Generalized Estimating Equation model because data with a Poisson distribution are non-normally distributed.

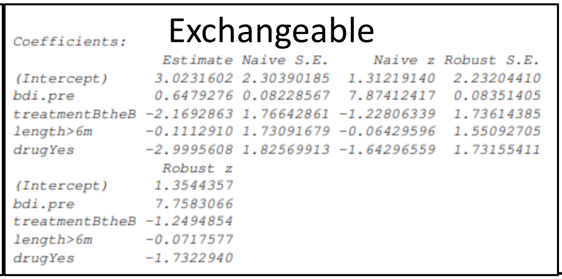
3. Below are summary results for two GEEs with different correlation structures. Which model is a more realistic fit to the data? Why?

The one with the exchangeable correlation matrix. There is very little difference between the Naïve and Robust S.E.



Independent

```
Coefficients:
                 Estimate Naive S.E.   Naive z Robust S.E.
(Intercept)     3.5686314  1.4833349  2.405816  2.26947617
bdi.pre         0.5818494  0.0563904 10.318235  0.09156455
treatmentBtheB -3.2372285  1.1295569 -2.865928  1.77459534
length>6m       1.4577182  1.1380277  1.280916  1.48255866
drugYes        -3.7412982  1.1766321 -3.179667  1.78271179
                 Robust z
(Intercept)     1.5724472
bdi.pre         6.3545274
treatmentBtheB -1.8242066
length>6m       0.9832449
drugYes        -2.0986557
```

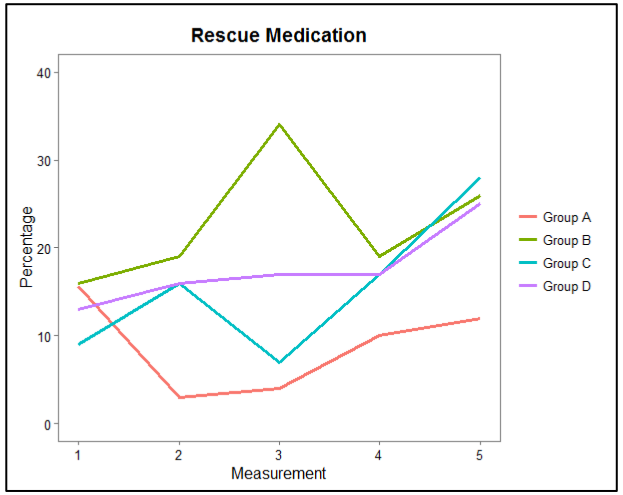


Exchangeable

```
Coefficients:
                 Estimate Naive S.E.     Naive z Robust S.E.
(Intercept)     3.0231602 2.30390185  1.31219140  2.23204410
bdi.pre         0.6479276 0.08228567  7.87412417  0.08351405
treatmentBtheB -2.1692863 1.76642861 -1.22806339  1.73614385
length>6m      -0.1112910 1.73091679 -0.06429596  1.55092705
drugYes        -2.9995608 1.82569913 -1.64296559  1.73155411
                 Robust z
(Intercept)     1.3544357
bdi.pre         7.7583066
treatmentBtheB -1.2494854
length>6m      -0.0717577
drugYes        -1.7322940
```

4. To the right is a spaghetti plot of the percent of patients who took rescue medication by group. Does there appear to be differences across group and time (measurement)?

Yes, the groups start around the same %, but diverge across time. By the final measurement, group A is lower than the other three, which are roughly the same.

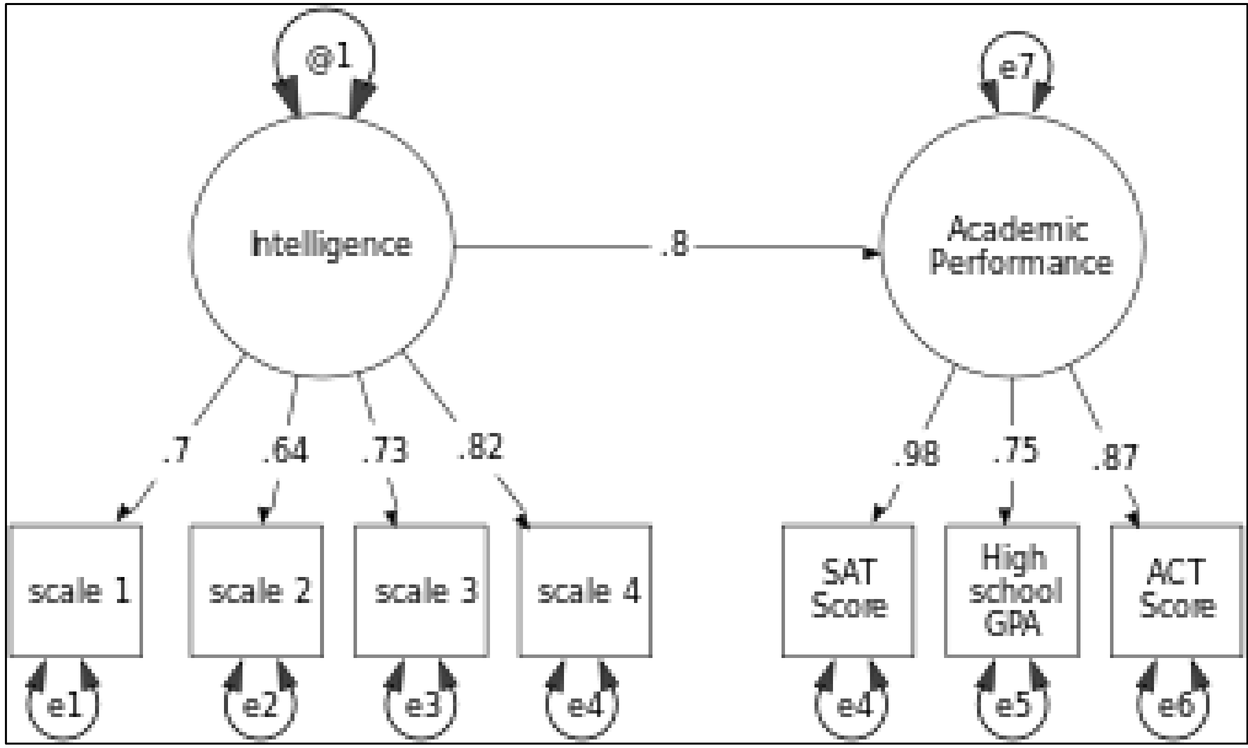


Structural Equation Modeling

Descriptions

- Multivariate statistical analysis-> factor analysis combined with multiple regression analysis
- Can be used to impute relationships between unobserved constructs (latent variables) from observable variables
- General approach
  - Model specification
  - Estimation of free parameters
  - Assessment of model and model fit
  - Model modification
  - Sample size and power
  - Interpretation and communication

- **Boxes: observed variables**
- **Circles: unobserved (latent) variables**
- **Arrows: paths**
  - **pointing: first variable affects the second (First -> Second)**
  - **small number is the value of constrained path coefficient**
  - **no number, then coefficient estimated from the data**
- **Curved, double headed paths: covariance (not otherwise assumed, like exogenous variables)**



An example structural equation model. Latent variables are drawn as circles. Manifest or measured variables are shown as squares. Residuals and variances are drawn as double headed arrows into an object. Note latent IQ variable fixed at 1 to provide scale to the model (Wikipedia)

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF
NORTH DAKOTA

Structural Equation Modeling

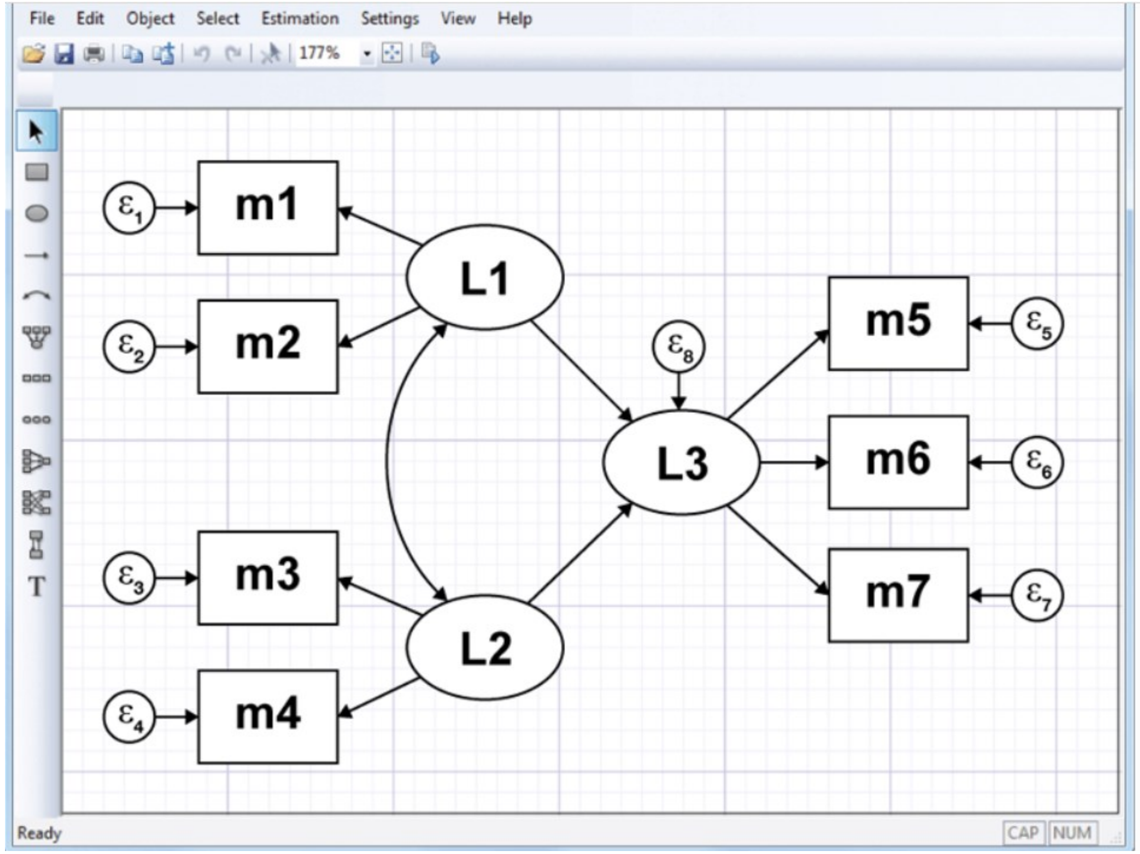
Formats

Command Line:

```
. sem (L1 -> m1 m2)
        (L2 -> m3 m4)
        (L3 <- L1 L2)
        (L3 -> m5 m6 m7)
```

```
. sem (m1 <- L1) (m2 <- L1) (L2 -> m3) (L2 -> m4)
(L3 -> m5) (L3 -> m6) (L3 -> m7) (L3 <- L1) (L3 <- L2)
cov(e.m1*e.m2) cov (e.L1*e.L2)
```

Graphically:



DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF
NORTH DAKOTA

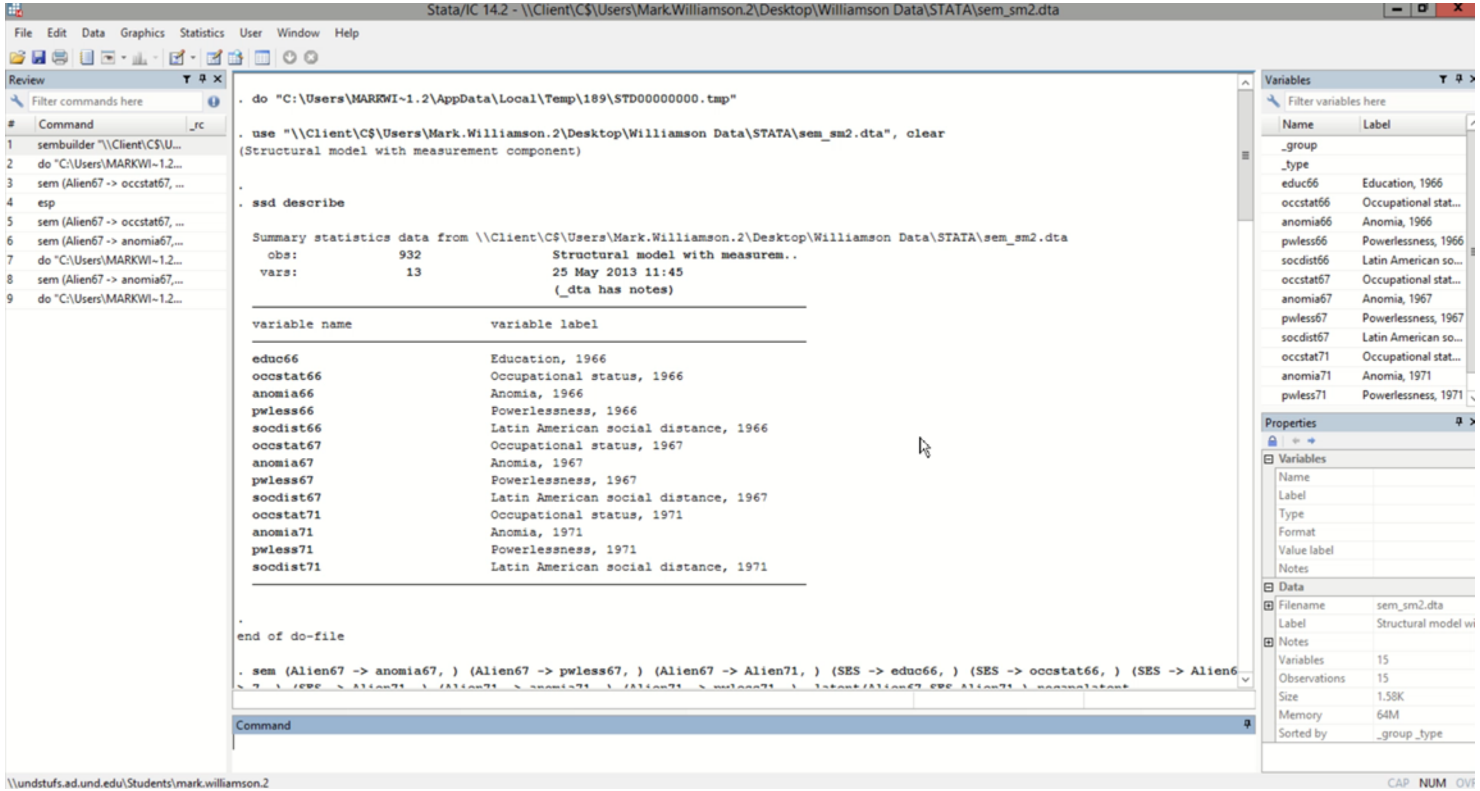
Structural Equation Modeling

Examples

Wheaton[1]

Fictional Data[2]

Affective/Cognitive Arousal[2]


Stata/IC 14.2 - \\Client\C$\Users\Mark.Williamson.2\Desktop\Williamson Data\STATA\sem_sm2.dta
File
Edit
Data
Graphics
Statistics
User
Window
Help
Review
Filter commands here
Variables
Filter variables here
Properties
Command
\\undstufs.ad.und.edu\Students\mark.williamson.2
CAP
NUM
OVR

#	Command	_rc
1	sembuilder "\\Client\C$\U...	
2	do "C:\Users\MARKWI~1.2...	
3	sem (Alien67 -> occstat67, ...	
4	esp	
5	sem (Alien67 -> occstat67, ...	
6	sem (Alien67 -> anomia67,...	
7	do "C:\Users\MARKWI~1.2...	
8	sem (Alien67 -> anomia67,...	
9	do "C:\Users\MARKWI~1.2...	

```
. do "C:\Users\MARKWI~1.2\AppData\Local\Temp\189\STD00000000.tmp"

. use "\\Client\C$\Users\Mark.Williamson.2\Desktop\Williamson Data\STATA\sem_sm2.dta", clear
(Structural model with measurement component)


.
. ssd describe

  Summary statistics data from \\Client\C$\Users\Mark.Williamson.2\Desktop\Williamson Data\STATA\sem_sm2.dta
   obs:            932                 Structural model with measurem..
  vars:             13                 25 May 2013 11:45
                                       (_dta has notes)
  ------------------------------------------------------------------
  variable name                variable label
  ------------------------------------------------------------------
  educ66                       Education, 1966
  occstat66                    Occupational status, 1966
  anomia66                     Anomia, 1966
  pwless66                     Powerlessness, 1966
  socdist66                    Latin American social distance, 1966
  occstat67                    Occupational status, 1967
  anomia67                     Anomia, 1967
  pwless67                     Powerlessness, 1967
  socdist67                    Latin American social distance, 1967
  occstat71                    Occupational status, 1971
  anomia71                     Anomia, 1971
  pwless71                     Powerlessness, 1971
  socdist71                    Latin American social distance, 1971
  ------------------------------------------------------------------


.
end of do-file

. sem (Alien67 -> anomia67, ) (Alien67 -> pwless67, ) (Alien67 -> Alien71, ) (SES -> educ66, ) (SES -> occstat66, ) (SES -> Alien6
> 7, ) (SES -> Alien71, ) (Alien71 -> anomia71, ) (Alien71 -> pwless71, ), latent(Alien67 SES Alien71 ) nocapslatent
```

Name	Label
_group	
_type	
educ66	Education, 1966
occstat66	Occupational stat...
anomia66	Anomia, 1966
pwless66	Powerlessness, 1966
socdist66	Latin American so...
occstat67	Occupational stat...
anomia67	Anomia, 1967
pwless67	Powerlessness, 1967
socdist67	Latin American so...
occstat71	Occupational stat...
anomia71	Anomia, 1971
pwless71	Powerlessness, 1971

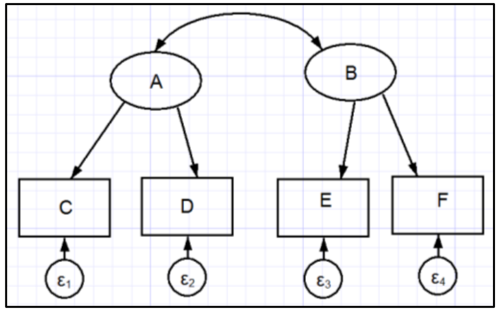
⊟ Variables	
Name	
Label	
Type	
Format	
Value label	
Notes	
⊟ Data	
⊞ Filename	sem_sm2.dta
Label	Structural model wi
⊞ Notes	
Variables	15
Observations	15
Size	1.58K
Memory	64M
Sorted by	_group _type

[1] https://www.stata.com/stata12/structural-equation-modeling/

[2] https://www.stata.com/manuals13/sem.pdf

Assessment 3

1. In the mock SEM diagram below, what type of variable does B represent? What about E?



2. Which of the following Stata commands would include the covariance between VarA and VarB?

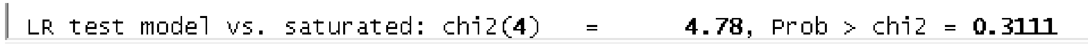
a) cov(e.VarA) cov(e.VarB)

b) cov(e.VarA*e.VarB)

c) cov(e.VarA, e.VarB)

c) cov(e.VarA e.VarB)

3. Looking at the fit of a SEM model returned the values below. Is the model a good fit? Why or why not?



```
LR test model vs. saturated: chi2(4)    =       4.78, Prob > chi2 = 0.3111
```

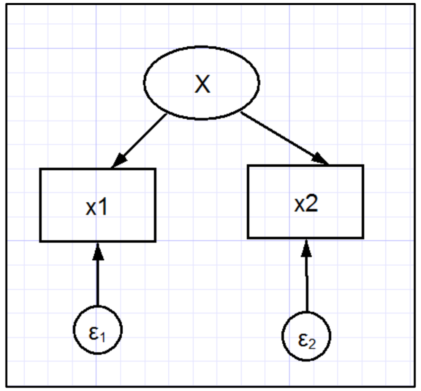
4. What is the proper line path for the SEM diagram to the right?

a) (x1 x2 <- X)

b) (x1 x2 -> X)

c) (X <- x1 x2)

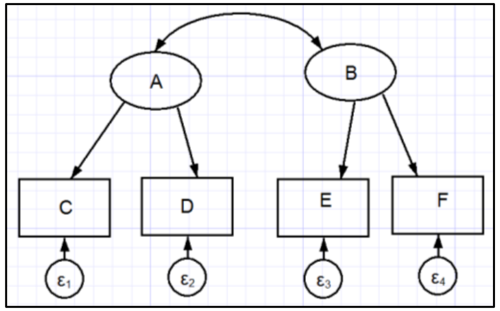
d) (X <- x1 <- x2)



Assessment 3

1. In the mock SEM diagram below, what type of variable does B represent? What about E?

B represents a latent, or unobserved variable (circle). E represents an observed variable (square).

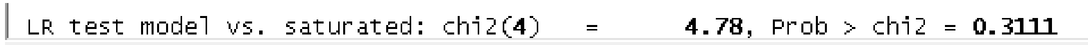


2. Which of the following Stata commands would include the covariance between VarA and VarB?

a) cov(e.VarA) cov(e.VarB) **b) cov(e.VarA*e.VarB)**

c) cov(e.VarA, e.VarB) c) cov(e.VarA e.VarB)

3. Looking at the fit of a SEM model returned the values below. Is the model a good fit? Why or why not?



```
LR test model vs. saturated: chi2(4)    =      4.78, Prob > chi2 = 0.3111
```

The test about if the model against a saturated model. Because the probability is not significant (>=0.05), the model is a good fit because there is no indication that adding more paths would improve the fit.

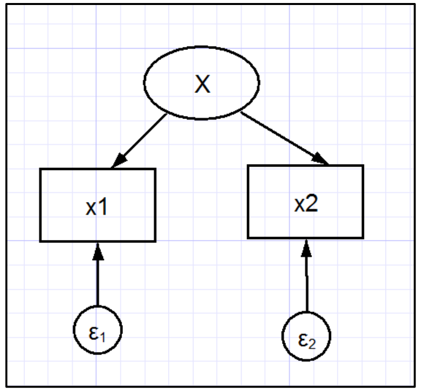
4. What is the proper line path for the SEM diagram to the right?

a) (x1 x2 <- X)

b) (x1 x2 -> X)

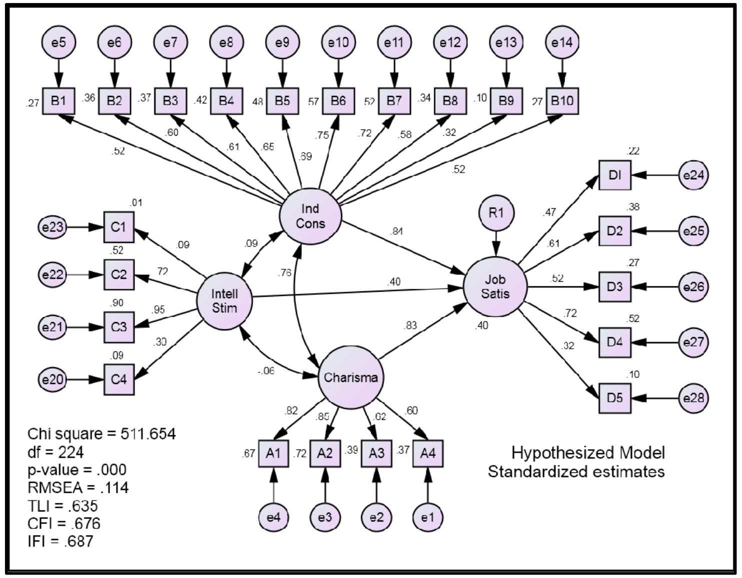
c) (X <- x1 x2)

d) (X <- x1 <- x2)



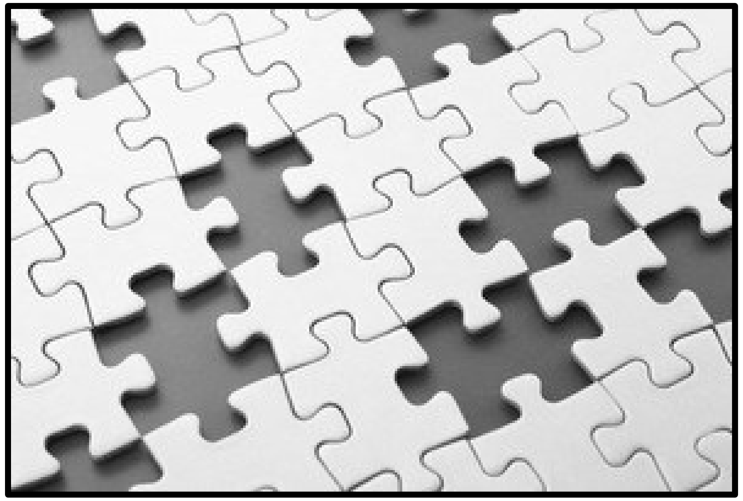
Caveats and Concerns

- More complex analyses come with more work and understanding
  - Multiple models, assumptions, and tests
- Data issues:
  - Restructuring
  - Reformatting
  - Missing data
- May need to try different software to get the job done



Chi square = 511.654
df = 224
p-value = .000
RMSEA = .114
TLI = .635
CFI = .676
IFI = .687

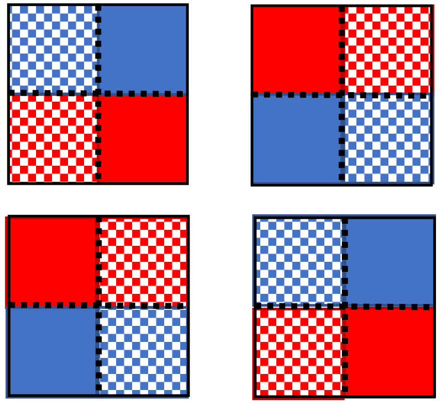
Hypothesized Model
Standardized estimates



Real World Examples

JONES, K. L., TODD, T. C., WALL-BEAM, J. L., COOLON, J. D., BLAIR, J. M., & HERMAN, M. A. (2006). Molecular approach for assessing responses of microbial-feeding nematodes to burning and chronic nitrogen enrichment in a native grassland. *15*(9), 2601-2609. doi:10.1111/j.1365-294X.2006.02971.x

Strip-Split



```
random block
       block*color
       block*shading;
```

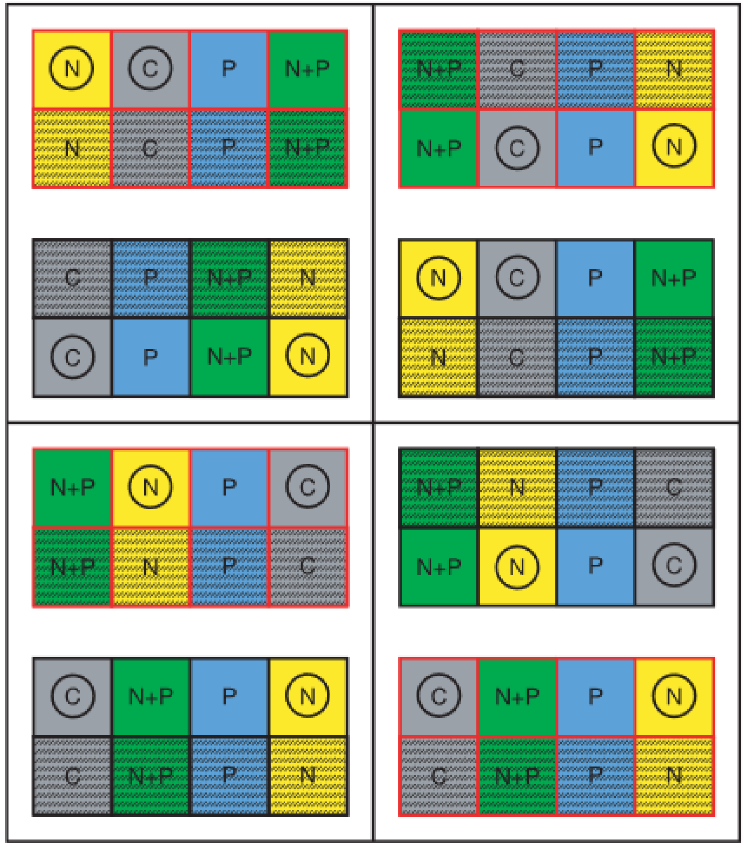
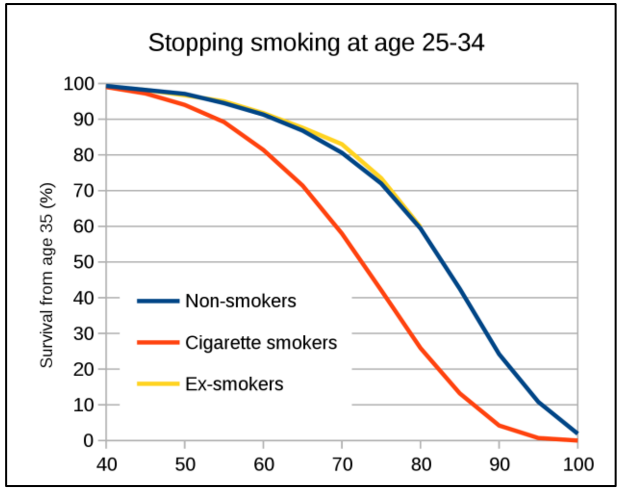
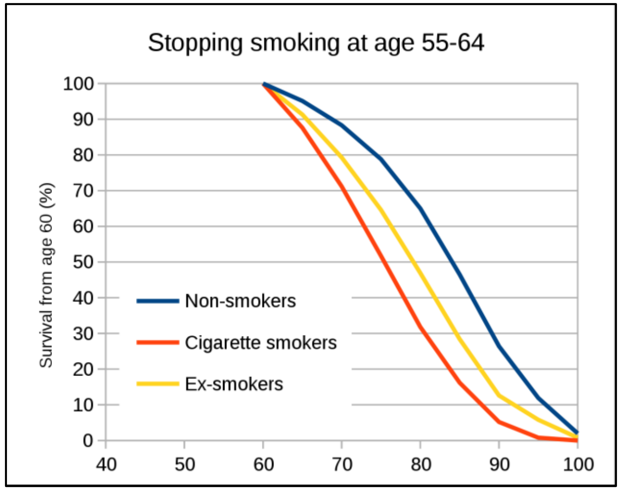


Fig. 1 The field experimental design included eight whole plots grouped into four blocks, with one plot per block that was burned annually (red outline) or left unburned (black outline). A split-strip plot design was obtained by mowing (hatched plots) or not mowing (open plots) one-half of each whole plot (i.e. the whole plots were split by mowing treatment) and using nutrient enrichment [nitrogen (N), phosphorous (P), both (N + P), or neither (C)] as a strip treatment applied perpendicular across the mowing treatments of each block. For this study, we sampled the 16 subplots (circled) that were not mowed, and had either nitrogen enrichment alone or had no nutrient addition. Thus, the four treatment combinations sampled were burned with and without nitrogen addition, and unburned with and without nitrogen addition.

Real World Examples

Doll, R., & Hill, A. B. (2004). The mortality of doctors in relation to their smoking habits: a preliminary report. 1954. *BMJ (Clinical research ed.), 328*(7455), 1529-1533. doi:10.1136/bmj.328.7455.1529





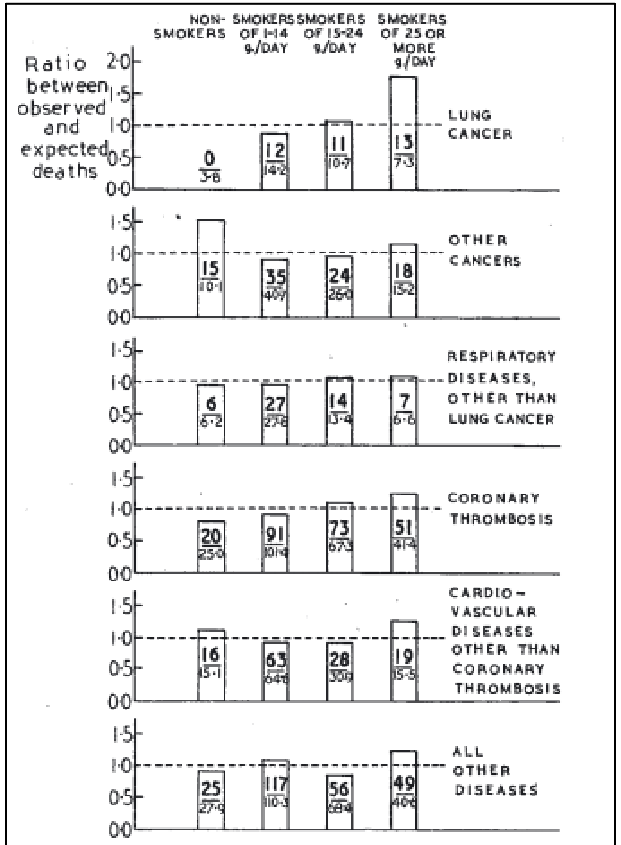


Chart showing variation in mortality with amount smoked. The ordinate shows the ratio between the number of deaths observed and the number expected (as entered in each column).

Real World Examples 🌎

Schwartz, G. G., & Klug, M. G. (2019). Thyroid Cancer Incidence Rates in North Dakota are Associated with Land and Water Use. *International Journal of Environmental Research and Public Health, 16*(20). doi:10.3390/ijerph16203805

Schwartz, G. G., Klug, M. G., & Rundquist, B. C. (2019). An exploration of colorectal cancer incidence rates in North Dakota, USA, via structural equation modeling. *International Journal of Colorectal Disease, 34*(9), 1571-1576. doi:10.1007/s00384-019-03352-9

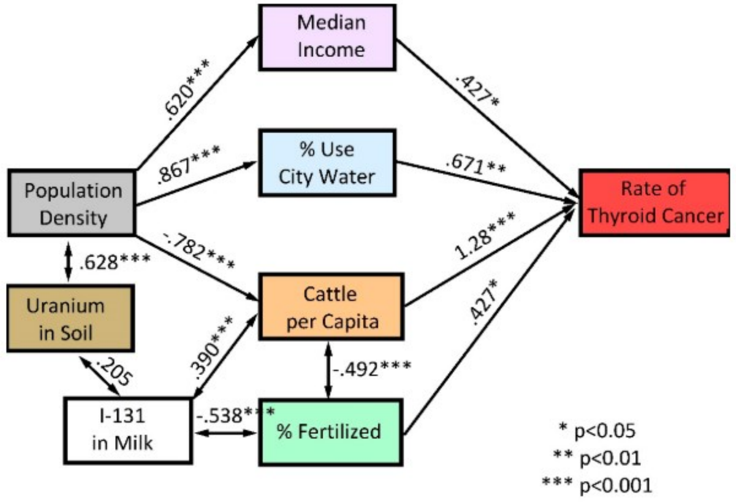
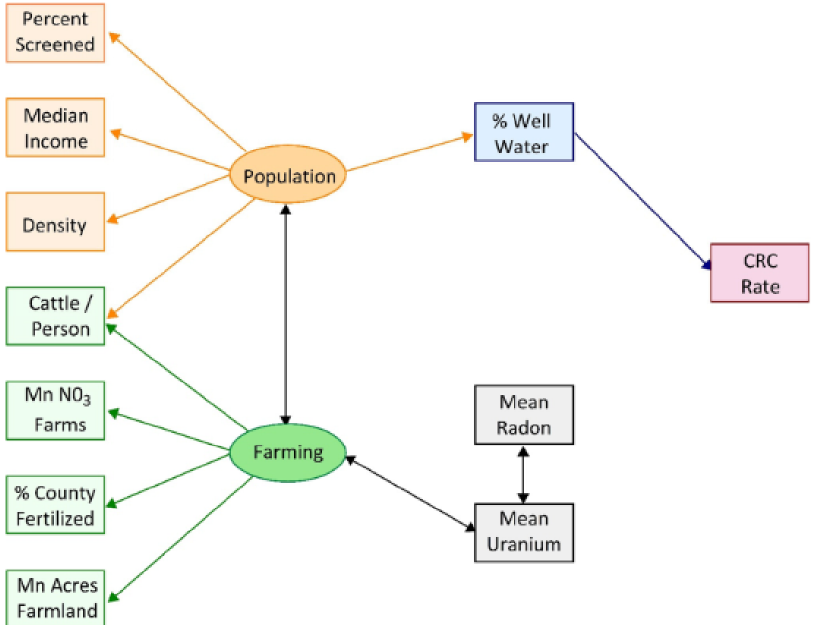


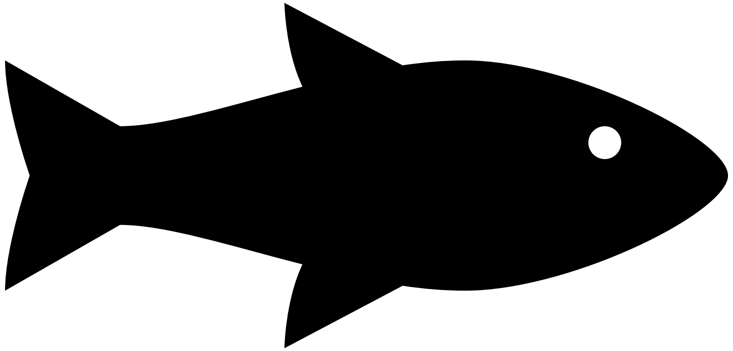
Figure 2. Structural Equation Model for county-specific thyroid cancer incidence rates in North Dakota. Uni-directional arrows indicate potential causal pathways. Bi-directional arrows indicate a co-varying relationship that is unlikely to be causal.

Fig. 3 Structural equation model of risk factors for CRC incidence among counties in ND. Measured variables are shown in rectangles, "latent variables" (i.e., variables that best represent a cluster of measured variables) are shown in oval outline. Uni-directional arrows indicate potential causal pathways; bi-directional arrows indicate a co-varying relationship that is unlikely to be causal. Mn = mean



Summary and Conclusion

- There are lots of advanced regression approaches
- Approach depends on data and questions asked
- Requires more work, understanding, and patience the more complex it is
- R, SAS, and STATA all have procedures for advanced approaches



Acknowledgements

- The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.
- For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications. ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)".***

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY