Regression Methods

GEE and Imputation

Samuel Alan Stewart, PhD

Medical Informatics
Department of Community Health & Epidemiology
Dalhousie University, Halifax, Canada
sam.stewart@dal.ca
@medInfProf

July 26, 2017

Poisson Regression Example


```
//read the data in
use https://stats.idre.ucla.edu/stat/stata/dae/
    poisson_sim, clear
```

- Perform basic exploration on the data (to better understand the variables)
- Build a simple Poisson regression model predicting the number of awards a student might win
  - You choose the predictors that you think are important

Simple Data Summary


```
sum num_awards math
tab prog
tab num_awards
tab prog, sum(num_awards)
tab prog num_awards
tab prog, sum(math)
```

Poisson Regression


```
poisson num_awards i.prog math, vce(robust) irr
//testing the program variable
test 2.prog 3.prog
//same as above command, but at variable level
testparm i.prog
//testing the math variable (same result as WALD test)
testparm math
```

Results

```
Iteration 0:   log pseudolikelihood = -182.75759
Iteration 1:   log pseudolikelihood = -182.75225
Iteration 2:   log pseudolikelihood = -182.75225

Poisson regression                              Number of obs     =        200
                                                Wald chi2(3)      =      80.15
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -182.75225               Pseudo R2         =     0.2118

------------------------------------------------------------------------------
             |               Robust
  num_awards |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        prog |
   academic  |   2.956065   .9514208     3.37   0.001     1.573083    5.554903
   vocation  |   1.447458   .5810418     0.92   0.357     .6590449    3.179049
             |
        math |   1.072672   .0112216     6.71   0.000     1.050902    1.094893
       _cons |   .0052626   .0034082    -8.10   0.000     .0014789    .0187265
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

Results

```
. test 2.prog 3.prog

 ( 1)  [num_awards]2.prog = 0
 ( 2)  [num_awards]3.prog = 0

           chi2(  2) =   14.76
         Prob > chi2 =    0.0006

. testparm i.prog

 ( 1)  [num_awards]2.prog = 0
 ( 2)  [num_awards]3.prog = 0

           chi2(  2) =   14.76
         Prob > chi2 =    0.0006

. testparm math

 ( 1)  [num_awards]math = 0

           chi2(  1) =   44.97
         Prob > chi2 =    0.0000
```

Results

- Is the model a significant fit?
- Which variables are significant fits? What is their p-value?

Results

- Is the model a significant fit?
  - Yes, since the `Wald chi2(3)` statistic is significant
- Which variables are significant fits? What is their p-value?

Results

- Is the model a significant fit?
  - Yes, since the `Wald chi2(3)` statistic is significant
- Which variables are significant fits? What is their p-value?
  - Prog is with a p-value of 0.0006

Results

- Is the model a significant fit?
  - Yes, since the `Wald chi2(3)` statistic is significant
- Which variables are significant fits? What is their p-value?
  - Prog is with a p-value of 0.0006
  - Math is with a p-value of < 0.0001

Results

- What is the effect of program?
- What is the effect of math?

Results

- What is the effect of program?
  - We expect someone in the academic stream to get 2.95 times more awards than someone in the general stream (95%CI: [1.6, 5.6])
- What is the effect of math?

Results

- What is the effect of program?
  - We expect someone in the academic stream to get 2.95 times more awards than someone in the general stream (95%CI: [1.6, 5.6])
  - There is little evidence that people in the vocational stream differ from the academic stream, with a HR of 1.44 (95% CI: [0.7,3.2])
- What is the effect of math?

Results

- What is the effect of program?
  - We expect someone in the academic stream to get 2.95 times more awards than someone in the general stream (95%CI: [1.6, 5.6])
  - There is little evidence that people in the vocational stream differ from the academic stream, with a HR of 1.44 (95% CI: [0.7,3.2])
- What is the effect of math?
  - For every point increase in math grade we expect a student to earn 1.07 times more awards (95%CI: [1.05, 1.09])

Modeling Process

1. Develop the model
   - This is where the type and components are defined
2. Estimate the coefficients
   - Usually done with (quasi-) Maximum Likelihood Estimation
3. Evaluate the model fit
   - Likelihood statistics, chi-square tests, Wald tests
4. Test the regression coefficients
   - Wald statistics for coefficients, likelihood tests for the variable
5. Test the regression assumptions
   - Most common assumption is independence

General Linear Modeling (GLM)

$$g(E(y)) = g(\mu) = X\beta + \epsilon$$

- $g(\mu)$ is some transformation of the expected value (i.e the mean) of the outcome such that the relationship is linear
- We looked at four different general linear models

Linear	regression is used when y is a continuous, somewhat normally distributed variable
Logistic	regression is used when Y is binary
Poisson	regression is used when Y is a counting variable
Negative Binomial	regression is used when the Poisson is over-dispersed

- There are many other forms of regression

Longitudinal Data

Longitudinal Data

- When we capture multiple values from a subject over time
  - Measuring weight every month for 12 months
  - Getting a patient's medication count every time they visit the ER
  - Recording a patient's smoking status whenever they show up to an AA meeting
- This can also relate to clustered data as well
  - Studying students that come from different schools
  - Studying patients seen in different ERs
- Why can't we use traditional methods?

Independence

- All models to this point have required independence, but what does independence mean?

Independence

- All models to this point have required independence, but what does independence mean?
- An observation is **dependent** if the value of the i^{th} observation is influenced by the $(i-1)^{th}$

Independence

- All models to this point have required independence, but what does independence mean?
- An observation is **dependent** if the value of the i^{th} observation is influenced by the $(i-1)^{th}$
  - This is true for responses or predictors

Independence

- All models to this point have required independence, but what does independence mean?
- An observation is **dependent** if the value of the i^{th} observation is influenced by the $(i-1)^{th}$
  - This is true for responses or predictors
- This is different from what we normally think of as a repeated measures or a split plot design, though both use the same subjects multiple times

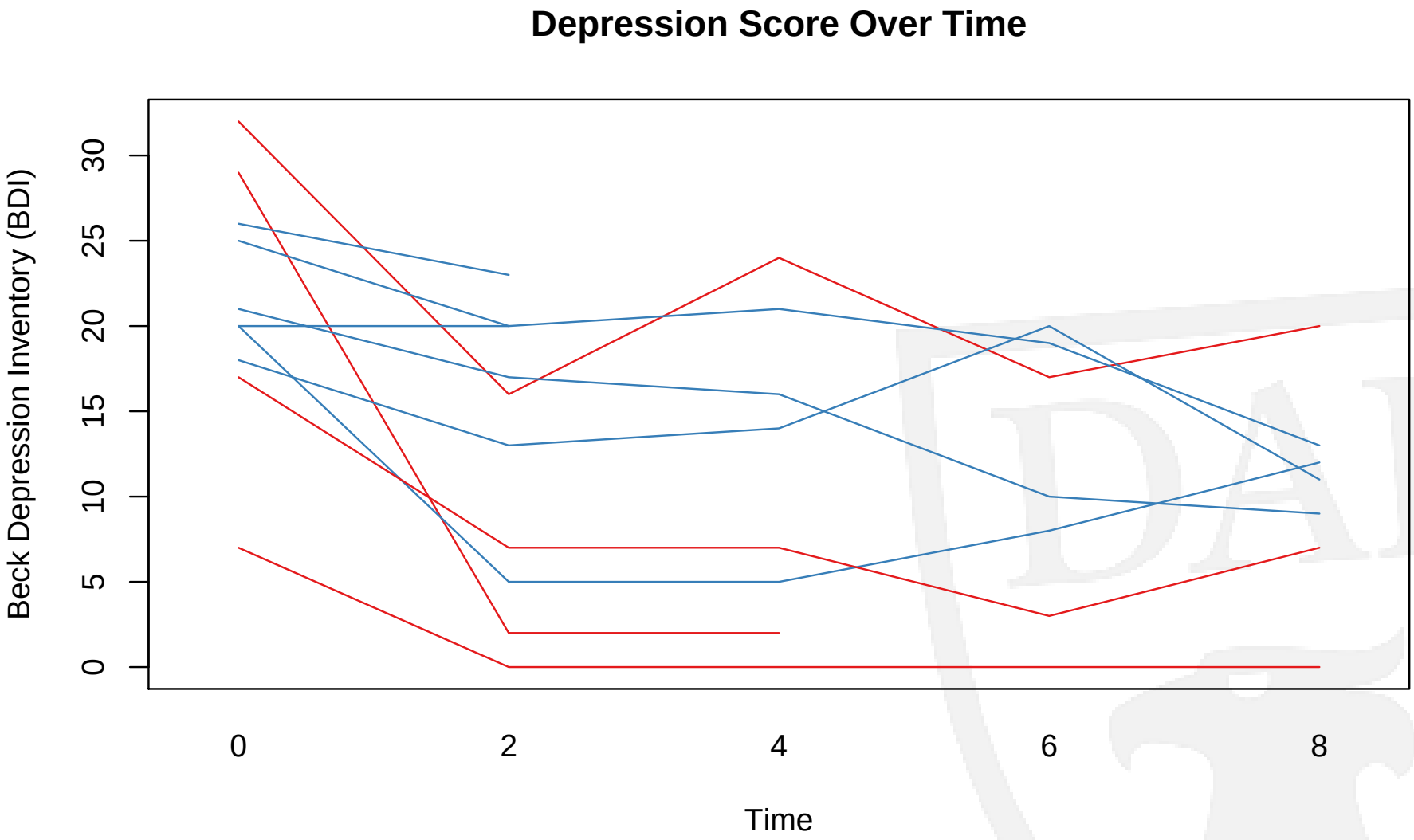
Repeated Measures

Pulse Measurements in beats/min

User	Trt 1	Trt 2	Trt 3
1	112	166	215
2	111	166	225
3	89	132	189
4	95	134	186
5	66	109	150
6	69	119	177
7	125	177	241
8	85	117	186
9	97	137	185
10	93	151	217
11	77	122	178
12	78	119	173
⋮			

Dependence in Depression Measurement

ID	drug	length	treatment	bdi.pre	bdi.2m	bdi.4m	bdi.6m	bdi.8m
1	No	>6m	TAU	29.00	2.00	2.00		
2	Yes	>6m	BtheB	32.00	16.00	24.00	17.00	20.00
3	Yes	<6m	TAU	25.00	20.00			
4	No	>6m	BtheB	21.00	17.00	16.00	10.00	9.00
5	Yes	>6m	BtheB	26.00	23.00			
6	Yes	<6m	BtheB	7.00	0.00	0.00	0.00	0.00
7	Yes	<6m	TAU	17.00	7.00	7.00	3.00	7.00
8	No	>6m	TAU	20.00	20.00	21.00	19.00	13.00
9	Yes	<6m	BtheB	18.00	13.00	14.00	20.00	11.00
10	Yes	>6m	BtheB	20.00	5.00	5.00	8.00	12.00


Depression Score Over Time
Beck Depression Inventory (BDI)
0
5
10
15
20
25
30
0
2
4
6
8
Time

Developing Longitudinal Models

- Our response variable is now Y_{ij} rather than Y_i, indicating that we have $i \in [1...n]$ subjects, each of whom have $j \in [1...n_i]$ observations
- We have k explanatory variables X_i which are patient-level measures
  - These are static variables, though time-varying covariates, X_{ij} are possible in GEE models
- In general what we want to do is estimate the equation $g(\mu) = \beta_0 + X_1\beta_1 + ... + X_k\beta_k + \epsilon$
  - How do we incorporate time?
  - How do we structure the error term, ϵ?

Building a Correlation Structure

GLM in Vector Notation

$$g(\mu) = X\beta + \epsilon \qquad \epsilon = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ 0 & 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}_{n \times n}$$

- In a GLM the correlation matrix is a diagonal matrix with constant value
- For longitudinal data this isn't the case, so we need to respecify the model

Building a Correlation Structure

GEE Model In Vector Notation

$$y_i = [y_{i1}, y_{i2}, y_{i3}, \ldots, y_{in_i}]^T \qquad y = [y_1, y_2, \ldots, y_n]^T$$

$$g(\mu) = X\beta + \epsilon \qquad \epsilon = \begin{bmatrix} \sigma^2 R & 0 & 0 & \ldots & 0 \\ 0 & \sigma^2 R & 0 & \ldots & 0 \\ 0 & 0 & \sigma^2 R & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \sigma^2 R \end{bmatrix}_{(\sum n_i)\times(\sum n_i)}$$

- The model is the same, but the covariance matrix is different
- R is the correlation matrix for the dependent measures, and its structure needs to be specified
- The solution is similar to a MLE approach

Potential Correlation Structures

Independent

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Autoregressive

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Exchangeable

$$\begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

Unstructured

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$$

1. Develop the Model

$$g(\mu) = X\beta + \epsilon \qquad R = ??$$

- Most of the model specification is the same as before
  - Identify the outcome type, identify the predictors
  - There's an additional step of deciding on the correlation structure, R_i

2. Estimate the Coefficients

DALHOUSIE UNIVERSITY

Penn State Stat 504: Analysis of Discrete Data

A quasi-likelihood estimate of β arise from maximization of normality-based loglikelihood without assuming that the response is normally distributed. In general, there are no closed-form solutions, so the GEE estimates are obtained by using an iterative algorithm, that is iterative quasi-scoring procedure.

GEE estimates of model parameters are valid even if the covariance is mis-specified *(because they depend on the first moment, e.g., mean).* **However, if the correlation structure is mis-specified, the standard errors are not good**, *and some adjustments based on the data(empirical adjustment) are needed to get more appropriate standard errors.*

3. Evaluate the Model Fit

- Since we did not use MLE, we don't have a deviance measure or a log-likelihood test
- QIC[1] is a metric that can be used to compare GEE models
  - We won't explore it's calculation here but you're welcome to.
  - It's designed to be like the AIC metric, so it already accounts for multiple predictors
- There are no tests for QIC, we use the model with the lowest QIC value

[1] Pan W. (2001) Akaike's Information Criterion in Generalized Estimating Equations. Biometrics 57: 120-125

4. Test the Regression Coefficients

- Since each coefficient has an estimate and a standard error we can perform z-tests
- We can calculate CIs for the estimates
- Using QIC and multiple models we can test variables as a whole (type III tests)

5. Test the Regression Assumptions

- Independence and Outliers
  - Independence beyond the dependence structure already built in
- A "reasonably close" correlation structure
  - One of the strengths of GEE is that, for large samples, the specification of the correlation structure is less important
  - "Large" is nebulous, context specific and never clear
  - Normal practice is to fit multiple correlation structures and compare estimates: if they're significantly different THEN you need to work out the correlation structure
- An adequate sample size

Beat the Blues Clinical Trial

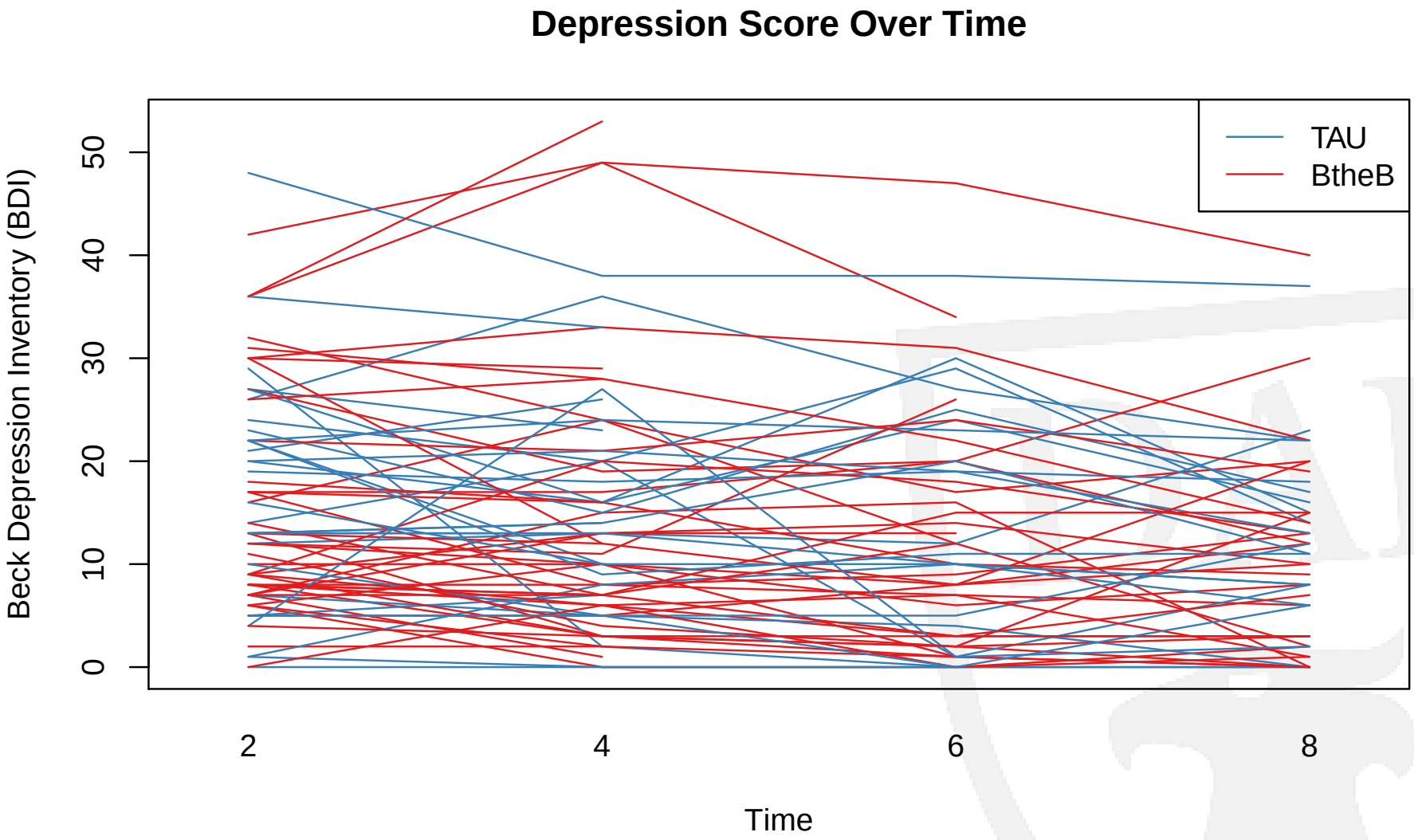
- Data are from the evaluation of an interactive multimedia program called "Beat the Blues"
- BtheB is a cognitive behavioural therapy delivered to depressed patients via a computer terminal.
- Patients with depression recruited in primary care were randomised to either the Beating the Blues program, or to Treatment as Usual (TAU)
- 100 patients recruited, 52 to the treatment
- Outcome variable: Beck depression inventory
- Predictor variables: centre, treatment, depression duration, presence of anti-depressants

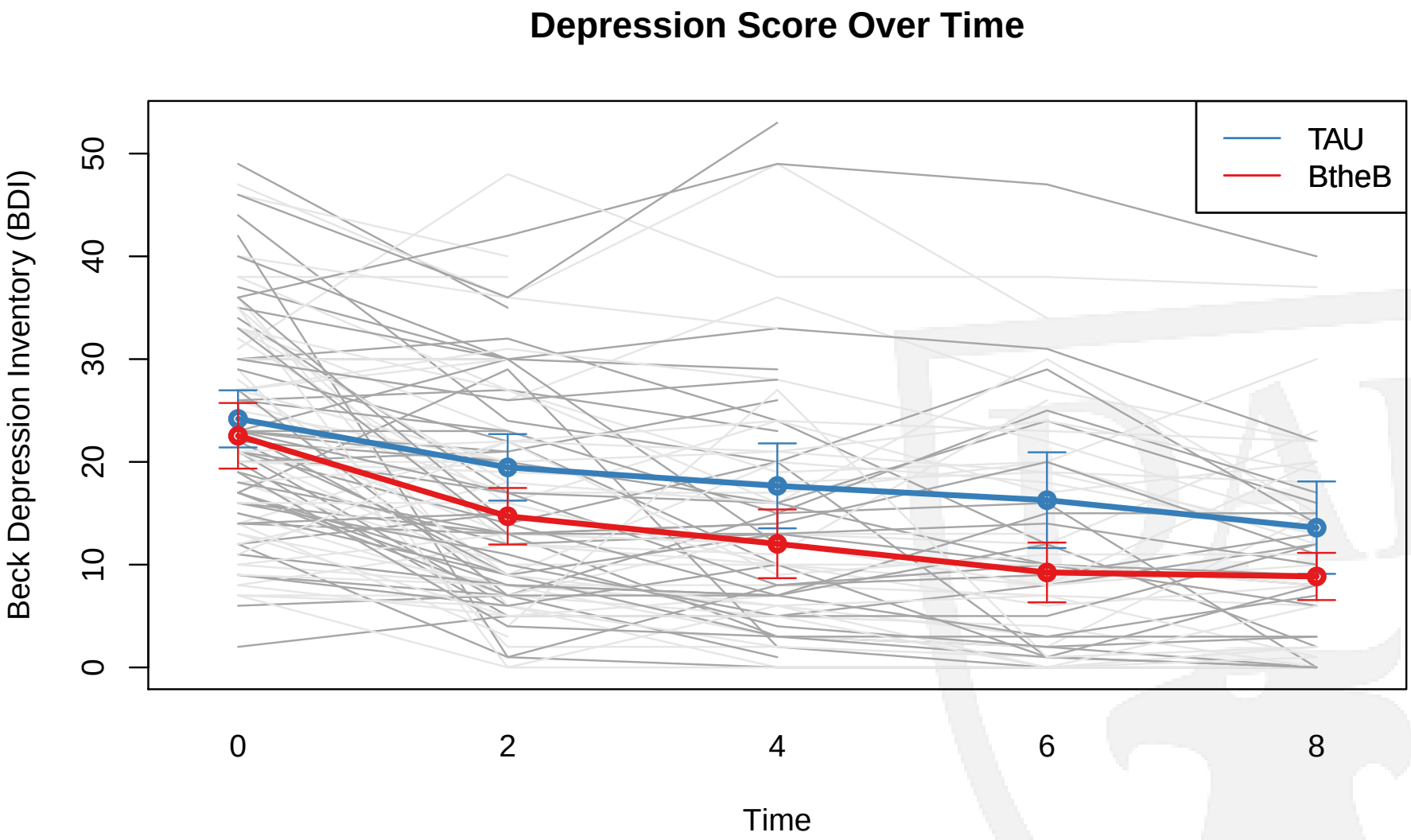
Summary

Variable	levels	n	TAU	BtheB
Drug	No	56	34 (0.607)	22 (0.393)
	Yes	44	14 (0.318)	30 (0.682)
Duration	<6m	49	23 (0.469)	26 (0.531)
	>6m	51	25 (0.49)	26 (0.51)

	BDI Averages (n, sd)				
	pre	2m	4m	6m	8m
TAU	24 (48,10)	19 (45,11)	18 (36,13)	16 (29,13)	14 (25,11)
BtheB	23 (52,12)	15 (52,10)	12 (37,10)	9 (29,8)	9 (27,6)

- Imbalance in drug use between groups
- Poor follow-up collection, many missing values


Depression Score Over Time
Beck Depression Inventory (BDI)
0
10
20
30
40
50
2
4
6
8
Time
TAU
BtheB


Depression Score Over Time
Beck Depression Inventory (BDI)
0
10
20
30
40
50
0
2
4
6
8
Time
TAU
BtheB

Building a GEE Model

- The hugely varying baseline BDI values should be taken into account
- We want to know the effect of the treatment, duration of depression and the presence of depression medication
- Need to consider the correlation structure

GEE Model

Independence Correlation Matrix

	Estimate	Robust S.E.	Robust z	pValue
(Intercept)	3.569	2.27	1.57	0.1158
bdi.pre	0.582	0.09	6.35	0.0000
treatmentBtheB	-3.237	1.77	-1.82	0.0681
length>6m	1.458	1.48	0.98	0.3255
drugYes	-3.741	1.78	-2.10	0.0358

Correlation Matrix

1.00	0.00	0.00	0.00
0.00	1.00	0.00	0.00
0.00	0.00	1.00	0.00
0.00	0.00	0.00	1.00

GEE Model

Exchangeable Correlation Matrix

	Estimate	Robust S.E.	Robust z	pValue
(Intercept)	3.023	2.23	1.35	0.1756
bdi.pre	0.648	0.08	7.76	0.0000
treatmentBtheB	-2.169	1.74	-1.25	0.2115
length>6m	-0.111	1.55	-0.07	0.9428
drugYes	-3.000	1.73	-1.73	0.0832

Correlation Matrix

1.00	0.68	0.68	0.68
0.68	1.00	0.68	0.68
0.68	0.68	1.00	0.68
0.68	0.68	0.68	1.00

GEE Model

Unstructured Correlation Matrix

	Estimate	Robust S.E.	Robust z	pValue
(Intercept)	3.248	2.25	1.44	0.1490
bdi.pre	0.624	0.09	7.30	0.0000
treatmentBtheB	-2.361	1.73	-1.36	0.1736
length>6m	0.259	1.55	0.17	0.8673
drugYes	-3.022	1.72	-1.75	0.0796

Correlation Matrix

1.00	0.64	0.52	0.42
0.64	1.00	0.57	0.46
0.52	0.57	1.00	0.59
0.42	0.46	0.59	1.00

Interpreting GEE Results

- The specification of the correlation makes a difference
  - This suggests that there is not a sufficient sample size (often a problem with GEE models)
- The unstructured and exchangeable look similar
  - The independence structure is rarely correct for temporal data
- Let's focus on the exchangeable model
  - Strongest effect on BDI was baseline BDI
  - Largest effect was presence of anti-depressants
  - Both anti-depressants and BtheB had expected effect, though neither are significant (sample size issue?)

In R

```
library(gee)
library(MuMIn)
data("BtheB", package = "HSAUR")
#converting the data to 'long' format
BtheB$subject <- factor(rownames(BtheB))
nobs <- nrow(BtheB)
BtheB_long <- reshape(BtheB, idvar = "subject",varying = c("bdi.2m", "bdi.4m", "bdi.6m",
     "bdi.8m"),direction = "long")
BtheB_long$time <- rep(c(2, 4, 6, 8), rep(nobs, 4))
#R function requires data be sorted by subject
BtheB_long =BtheB_long[order(BtheB_long$subject,BtheB_long$time),]
```

In R

```
mod01 =  gee(bdi ~ bdi.pre + treatment + length + drug, data = BtheB_long, id = subject,
     family = gaussian, corstr = "independence")
mod02 =  gee(bdi ~ bdi.pre + treatment + length + drug, data = BtheB_long, id = subject,
     family = gaussian, corstr = "exchangeable")
mod03 =  gee(bdi ~ bdi.pre + treatment + length + drug, data = BtheB_long, id = subject,
     family = gaussian, corstr = "unstructured")
mod03a =  gee(bdi ~ bdi.pre + treatment * drug + length, data = BtheB_long, id = subject,
      family = gaussian, corstr = "unstructured")
summary(mod01)
summary(mod02)
summary(mod03)
summary(mod03a)
QIC(mod01,mod02,mod03)
```

In STATA

- `xtgee` is the command to run a GEE in STATA
- Need `xtset` to tell STATA how to identify the subjects and the time variable
- Need to make sure to specify `robust` estimators
- Data is available in two data files: *BtheB.csv* and *BtheB_long.csv*
  - There are ways to transform the wide data into long in STATA if you so choose

Data Processing


```
import delimited "C:\Users\sstewar2\Documents\Teaching
    \Grad Students\RegressionMethodsCHE\BtheB.csv"

//Simple data summaries
tab drug
tab length
tab treatment
sum bdipre

tab drug treatment, row col
tab length treatment, row col
tab treatment, sum(bdipre)
```

Data Processing


```
//Importing LONG data, needed for GEE model
drop _all
import delimited "C:\Users\sstewar2\Documents\Teaching
    \Grad Students\RegressionMethodsCHE\BtheB_long.csv
    "

gen bdiValue = real(bdi)
encode drug, gen(drugFactor)
encode length, gen(lengthFactor)
encode treatment, gen(treatmentFactor)

//more data summaries
tab time treatment, sum(bdiValue)
```

Data Processing


```
//GEE in STATA with xtgee
xtset subject time

xtgee bdiValue bdipre drugFactor lengthFactor
    treatmentFactor, family(gaussian) cor(independent)
     robust
xtcorr

xtgee bdiValue bdipre drugFactor lengthFactor
    treatmentFactor, family(gaussian) cor(exchangeable
    ) robust
xtcorr

xtgee bdiValue bdipre drugFactor lengthFactor
    treatmentFactor, family(gaussian) cor(unstructured
    ) robust
xtcorr
```

Respiratory

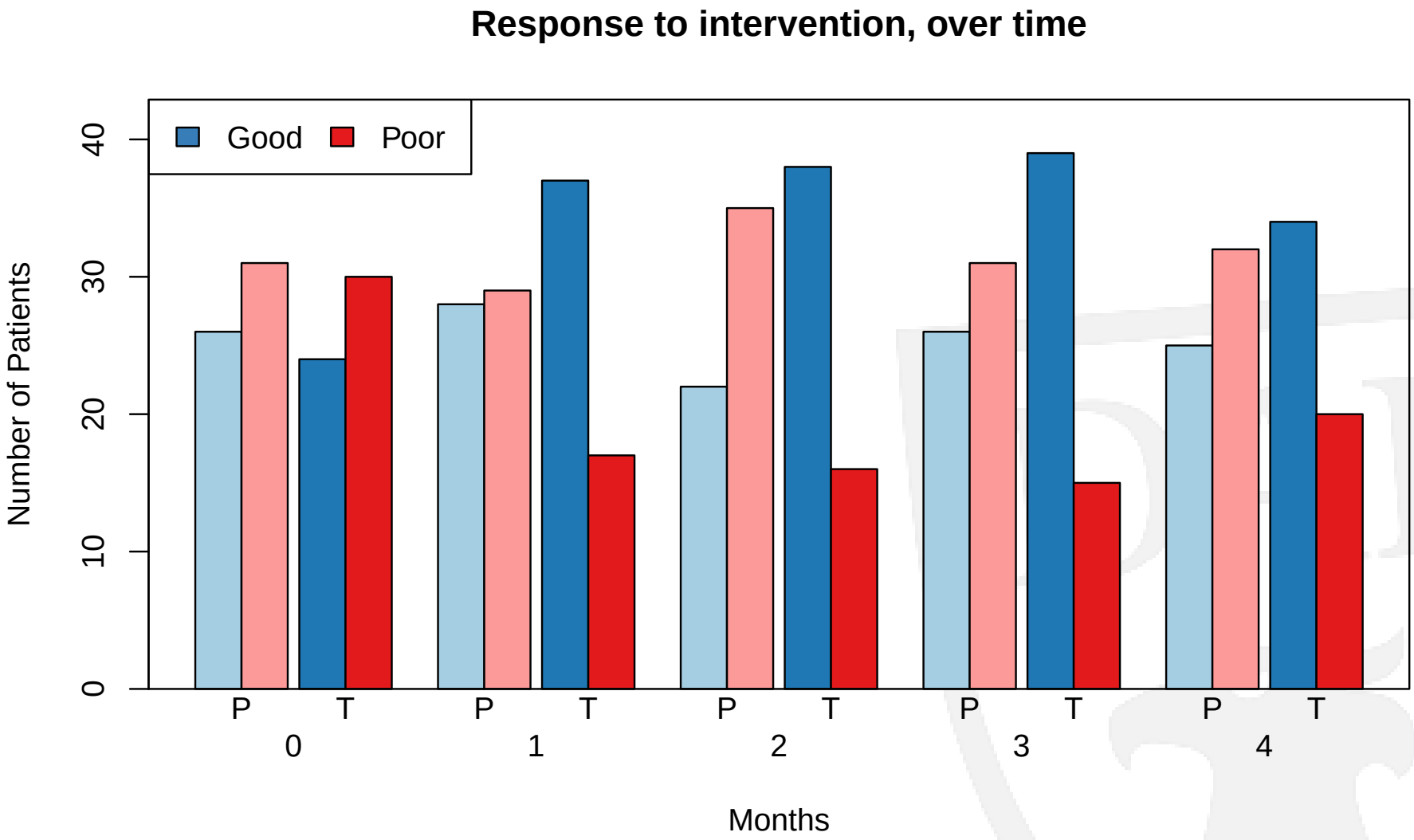
- Data from a multi-center study of some treatment
- 111 participants (54 treatment) each observed at 5 time-points (baseline + 4 follow-ups)
- Outcome variable: binary, respiratory status as **poor** or **good**
- Predictors: centre, treatment, gender and age

Summary

Variable	levels	n	placebo	treatment
Centre	1	56	29 (0.518)	27 (0.482)
	2	55	28 (0.509)	27 (0.491)
Gender	female	88	40 (0.455)	48 (0.545)
	male	23	17 (0.739)	6 (0.261)
Age		33.3 (14)	33.6 (13)	32.9 (14)

Proportion of Subjects with "Good" Status per visit

	0	1	2	3	4
placebo	0.46 (57)	0.49 (57)	0.39 (57)	0.46 (57)	0.44 (57)
treatment	0.44 (54)	0.69 (54)	0.7 (54)	0.72 (54)	0.63 (54)


Response to intervention, over time
Good
Poor
Number of Patients
0
10
20
30
40
P
T
0
1
2
3
4
Months

GEE Model

Independence Correlation Matrix

	Estimate	Robust S.E.	Robust z	pValue
(Intercept)	-0.900	0.46	-1.96	0.0505
centre2	0.672	0.36	1.88	0.0598
treatmenttreatment	1.299	0.35	3.70	0.0002
gendermale	0.119	0.44	0.27	0.7879
age	-0.018	0.01	-1.40	0.1624
baselinegood	1.882	0.35	5.38	0.0000

Correlation Matrix

1.00	0.00	0.00	0.00
0.00	1.00	0.00	0.00
0.00	0.00	1.00	0.00
0.00	0.00	0.00	1.00

GEE Model

Exchangeable Correlation Matrix

	Estimate	Robust S.E.	Robust z	pValue
(Intercept)	-0.900	0.46	-1.96	0.0505
centre2	0.672	0.36	1.88	0.0598
treatmenttreatment	1.299	0.35	3.70	0.0002
gendermale	0.119	0.44	0.27	0.7879
age	-0.018	0.01	-1.40	0.1624
baselinegood	1.882	0.35	5.38	0.0000

Correlation Matrix

1.00	0.34	0.34	0.34
0.34	1.00	0.34	0.34
0.34	0.34	1.00	0.34
0.34	0.34	0.34	1.00

GEE Model

Unstructured Correlation Matrix

	Estimate	Robust S.E.	Robust z	pValue
(Intercept)	-0.931	0.46	-2.02	0.0435
centre2	0.673	0.35	1.90	0.0579
treatmenttreatment	1.279	0.35	3.66	0.0003
gendermale	0.095	0.44	0.21	0.8310
age	-0.017	0.01	-1.31	0.1906
baselinegood	1.935	0.35	5.56	0.0000

Correlation Matrix

1.00	0.32	0.21	0.30
0.32	1.00	0.43	0.36
0.21	0.43	1.00	0.39
0.30	0.36	0.39	1.00

GEE Model

AR-1 Correlation Matrix

	Estimate	Robust S.E.	Robust z	pValue
(Intercept)	-0.963	0.46	-2.09	0.0368
centre2	0.743	0.36	2.08	0.0371
treatmenttreatment	1.247	0.35	3.54	0.0004
gendermale	0.113	0.45	0.25	0.8011
age	-0.017	0.01	-1.31	0.1907
baselinegood	1.911	0.35	5.46	0.0000

Correlation Matrix

1.00	0.39	0.15	0.06
0.39	1.00	0.39	0.15
0.15	0.39	1.00	0.39
0.06	0.15	0.39	1.00

In STATA


```
//GEE example from the R library HSAUR2
drop _all
import delimited "C:\Users\sstewar2\Documents\Teaching
    \Grad Students\RegressionMethodsCHE\respiratory.
    csv"

generate centre2 = centre>1
drop centre
rename centre2 centre
encode treatment, gen(t2)
drop treatment
rename t2 treatment
encode gender, gen (g2)
drop gender
rename g2 gender

generate nStatus = status=="good"
generate nBaseline = baseline=="good"
```

In STATA

```
//Summary Statistics
tab gender treatment if month==1
tab centre treatment if month==1
tab treatment if month==1,sum(age)
tab treatment if month==1,sum(nBaseline) nost

tab treatment month, sum(nStatus) nost
```

In STATA

```
//GEE in STATA with xtgee
xtset subject month
xtgee nStatus centre treatment nBaseline age gender,
    family(binomial) cor(independent) robust
xtcorr
xtgee nStatus centre treatment nBaseline age gender,
    family(binomial) cor(exchangeable) robust
xtcorr
xtgee nStatus centre treatment nBaseline age gender,
    family(binomial) cor(unstructured) robust
xtcorr
xtgee nStatus centre treatment nBaseline age gender,
    family(binomial) cor(ar) robust
xtcorr
```

Results

- More stable estimates (since the sample size is a bit better)
- The correlation estimates are interesting: $\rho \approx 0.35$ seems to be the overall estimate
- AR-1 model makes the most sense to me, so I'll go with that
- Treatment and baseline status are the largest and most significant factors
  - What do those coefficients mean?

Results

	Lower CI	OR	Upper CI
centre 2 vs 1	1.13	2.102	3.92
treatment vs control	1.88	3.481	6.44
Male vs Female	0.52	1.120	2.41
Age (1 unit decrease)	0.99	1.017	1.04
Good vs Poor Baseline	3.61	6.763	12.66

- Treatment patients odds of good response are 3.5 times higher
- Patients that start with "Good" response have odds 6.8 times higher to maintain "Good" response
- Centre 2 patients faired slightly better

Schizophrenia Data

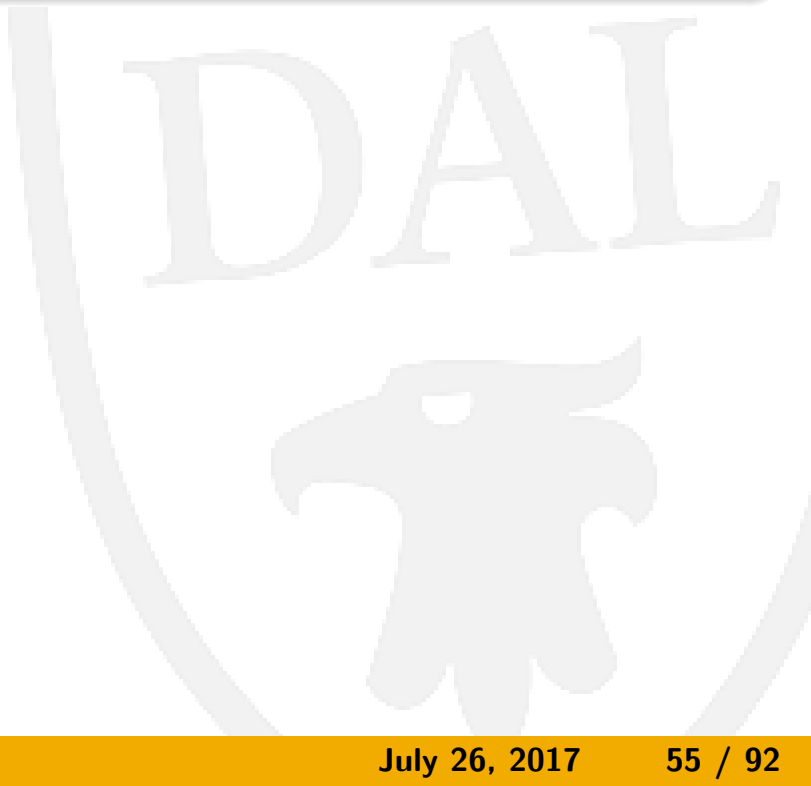
- A RT where 312 patients received drug therapy, 101 received placebo
- Measurements at 0, 1, 3, and 6 weeks (with some missing values)
- Outcome is severity of illness (7 point scale, 7=extremely ill)

In STATA

```
drop _all
copy https://onlinecourses.science.psu.edu/stat504/
    sites/onlinecourses.science.psu.edu.stat504/files/
    lesson09/schiz.dat schiz.dat
import delimited schiz.dat, delim(space) varnames(
    nonames)

rename v1 id
rename v2 group
rename v3 week
rename v4 severity
gen sqrtweek = sqrt(week)
```

Imputation



Imputation

- Need to deal with missing values in a dataset
- We'll look at the following missing-data methods:
  - Discarding
  - Specific-methods
  - Mean and Random imputation
  - Simple/random regression
  - Multiple Matching
  - Multiple Imputation

Missingness

Four Types of Missing Data

Missing Completely At Random (MCAR) When the value missing is completely uninfluenced by anything, and happens due to completely random nature. *Rarely happens*

Missing At Random (MAR) The mechanism driving the missingness can be entirely captured within the other variables in the dataset

Missing on Unobserved Values The mechanism driving the missingness is at least partially due to variables outside the scope of the project

Missing on Value The mechanism driving the missing value is the missing value itself

- Often the last two are grouped together as **Missing Not at Random (MNAR)**, though they are theoretically different

Four Types of Missing Data

Examples

Missing Completely At Random Some surveys were distributed with the question about depression missing from a 10 question survey

Missing At Random Men are less likely to answer questions about depression, regardless of their depression themselves

Missing on Unobserved Values People are less likely to answer questions about depression based on a variable called "belief in medicine", and this variable was not captured in the survey

Missing on Value Men are less likely to answer questions about depression if they suffer from depression

Dealing with Missing Data

Simplest Approach

Missing Completely At Random Can drop cases with missing values

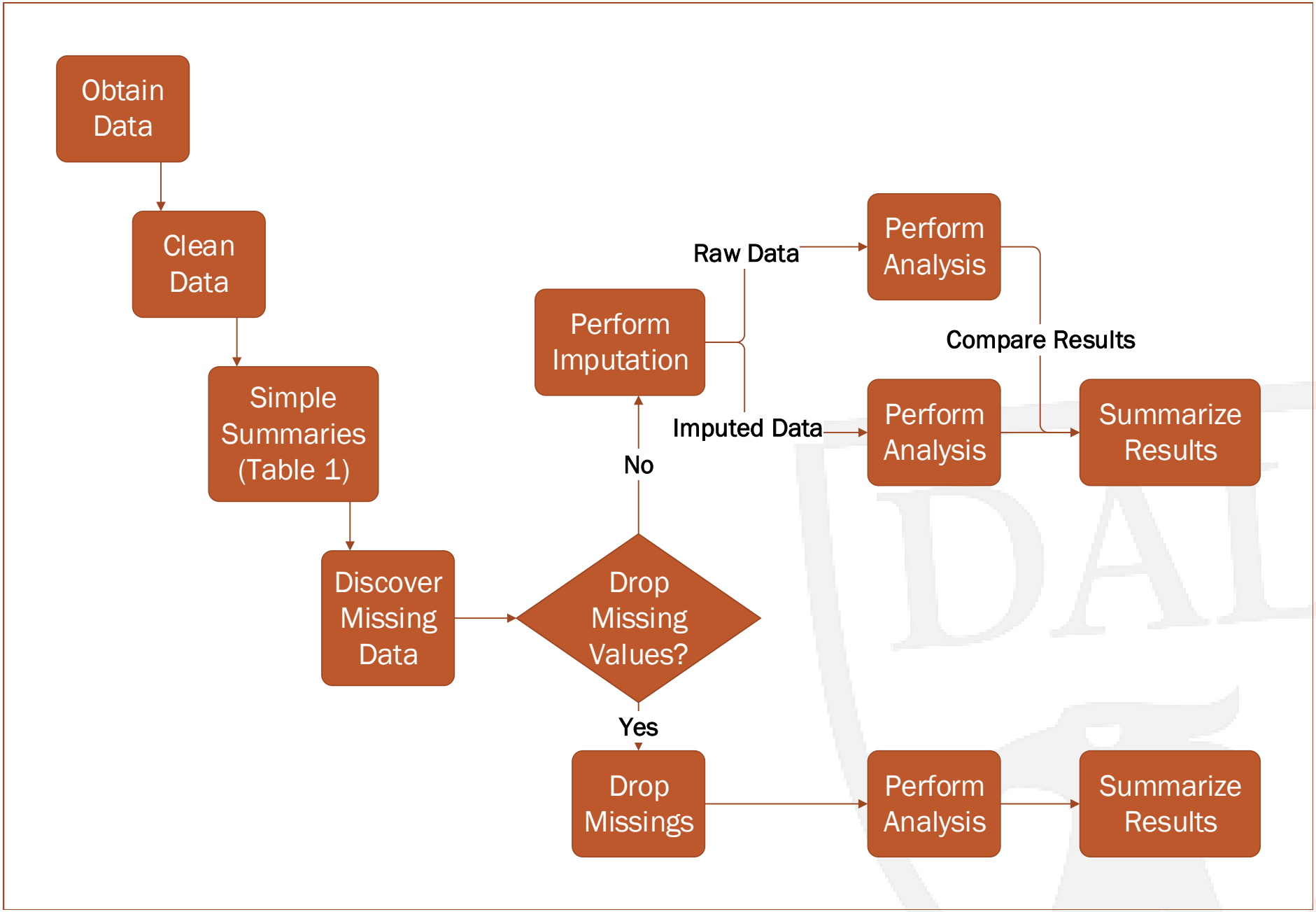
Missing At Random Can drop cases with missing values if the driving mechanisms are included

Missing on Unobserved Values Need to model the missingness to avoid biasing your results

Missing on Value Need to model the missingness to avoid biasing your results

How and When to Impute

- You never want to throw away data if you can avoid it
- It's always easiest to NOT impute
- If you do impute
  - Impute as late as possible
  - Make it clear that you imputed
  - Perform a sensitivity analysis (i.e compare to the non-imputed version)


Obtain Data
Clean Data
Simple Summaries (Table 1)
Discover Missing Data
Drop Missing Values?
No
Yes
Perform Imputation
Raw Data
Imputed Data
Perform Analysis
Perform Analysis
Compare Results
Summarize Results
Drop Missings
Perform Analysis
Summarize Results

Imputation Methods

- All imputation causes an under-estimation of variance
  - We should be increasing our variance estimates after imputation, since we're less certain about our data
  - Instead we decrease the variance slightly since we're adding more data to the analysis
- We'll look first a context specific methods for imputation, then move onto general mathematical approaches
- We'll start with notation

Imputation Notation

Really for This Lecture Only

- X is an $n \times k$ matrix of complete predictors
- Y is an $n \times c$ matrix of predictors with missing values

$Y =$

Y_1	Y_2	Y_3	Y_4
1			N
	B	7	Y
3	D	9	N
	D	9	Y
	C		
5	B	9	N
	B	9	Y
1	A		Y
5			Y
	B	9	Y

$X =$

X_1	X_2	X_3	X_4	X_5	X_6
Y	6	W	33	0	M
Y	6	Z	22	0	F
Y	5	W	22	1	M
N	4	Y	11	0	M
Y	3	Z	33	1	F
N	5	W	22	0	M
N	5	W	11	0	F
Y	9	X	11	1	M
N	2	W	22	1	F
N	6	Y	11	0	M

Context-Specific Methods

Last-value Carried Forward/Backward For studies that have a pre-post intervention, if one of the values is missing we can carry the last value forward, essentially assuming no effect

Related Observation Sometimes a value can be inferred from another value. If "Salary" is missing but they reported 0-months of work in a separate value, we can infer that their salary is probably 0. Very niche, but useful to explore in large datasets

Missing As A Level For categorical data we SHOULD study missing as it's own category (at least in table 1). People that won't report their salary are worth studying as a separate group

- Important to differentiate between didn't answer, chose not to answer, not applicable, ...

Mean Imputation

- For continuous missing variables, use the mean of the value as the missing value
- For categorical values, can use the mode (most common)
- Has many problems
  - Underestimates the SD
  - Pulls estimates towards the mean
- Is useful as a first step, or filling in when you have few missing values

Mean Imputation

$Y =$

Y_1	Y_2	Y_3	Y_4
1			N
	B	7	Y
3	D	9	N
	D	9	Y
	C		
5	B	9	N
	B	9	Y
1	A		Y
5			Y
	B	9	Y

$Y_{MeanImp} =$

Y_1	Y_2	Y_3	Y_4
1	**B**	**9**	N
3	B	7	Y
3	D	9	N
3	D	9	Y
3	C	**9**	**Y**
5	B	9	N
3	B	9	Y
1	A	**9**	Y
5	**B**	**9**	Y
3	B	9	Y

Random Imputation

- Fill in missing values with a random sample of the values in the variable
- Approach REALLY doesn't make sense, but it's a good first approach for later methods
- Less biasing than mean imputation

Random Imputation

$Y =$

Y_1	Y_2	Y_3	Y_4
1			N
	B	7	Y
3	D	9	N
	D	9	Y
	C		
5	B	9	N
	B	9	Y
1	A		Y
5			Y
	B	9	Y

$Y_{RandomImp} =$

Y_1	Y_2	Y_3	Y_4
1	**B**	**9**	N
3	B	7	Y
3	D	9	N
1	D	9	Y
5	C	**9**	**Y**
5	B	9	N
1	B	9	Y
1	A	**7**	Y
5	**B**	**9**	Y
1	B	9	Y

Regression Imputation

- Build a regression model predicting a single Y_i missing value with all the available X values
- Works well for data that is missing at random
- Requires a relationship to exist
- Should consider adding the error from the regression back into the model
- Can get complicated for non-binary categorical variables

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

$$Y_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

$$Y_4 : logit(p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

Regression Imputation

There are actually 100 more observations in each matrix

$Y =$

Y_1	Y_2	Y_3	Y_4
1			N
	B	7	Y
3	D	9	N
	D	9	Y
	C		
5	B	9	N
	B	9	Y
1	A		Y
5			Y
	B	9	Y

$Y_{RegressImp} =$

Y_1	Y_2	Y_3	Y_4
1.0		**6.4**	N
3.3	B	7.0	Y
3.0	D	9.0	N
3.2	D	9.0	Y
3.2	C	**6.9**	**Y**
5.0	B	9.0	N
2.7	B	9.0	Y
1.0	A	**7.3**	Y
5.0		**6.9**	Y
3.1	B	9.0	Y

An Aside on "Hot Deck"[2]

The History of Hot Deck

A once-common method of imputation where a missing value was imputed from a randomly selected similar record. The term "hot deck" dates back to the storage of data on punched cards, and indicates that the information donors come from the same dataset as the recipients. The stack of cards was "hot" because it was currently being processed.

- Traditional "Hot-deck" imputation is just random imputation
- Is now applied more liberally: any method that uses a different observation from the same variable can be called "hot-deck" imputation
- We're going to look at a way to pick the replacement observation better than "randomly"

Matching Imputation[3]

- For each subject with a missing value, find the subject "most-similar" to them in terms of their X values, and use their value
- Some issues that arise in matching:
  - How to break ties if multiple users "match"
  - How to measure "similar" across different variable types
  - Weighting certain variables to be more influential in the matching

[3] Cranmer, S.J. and Gill, J.M. (2013) We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data. British Journal of Political Science 43:2 (425-449)

Hot Deck Imputation

There are actually 100 more observations in each matrix

$Y =$

Y_1	Y_2	Y_3	Y_4
1			N
	B	7	Y
3	D	9	N
	D	9	Y
	C		
5	B	9	N
	B	9	Y
1	A		Y
5			Y
	B	9	Y

$Y_{HotDeckImp} =$

Y_1	Y_2	Y_3	Y_4
1	**C**	**1.0**	N
2	B	7.0	Y
3	D	9.0	N
2	D	9.0	Y
2	C	**5.0**	**Y**
5	B	9.0	N
1	B	9.0	Y
1	A	**5.0**	Y
5	**B**	**7.0**	Y
2	B	9.0	Y

Multiple Imputation[4]

- Multiple Imputation is MORE than an imputation strategy, it is an analytic approach
- Rather than trying to correct the imputation bias, let's try to average it out
  - Impute each missing value M times, creating M datasets
  - For each dataset, produce the model estimates (coefficients, OR, whatever your summary statistic is)
  - Combine the results of the M different summary statistics into a single measure
- Each imputation must be different in some way

[4] Sterne JA et al. (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009 Jun 29;338:b2393. doi: 10.1136/bmj.b2393.

Imputation in R

```
set.seed(11)
Y = data.frame(Y1=sample(c(NA,1:5),10,replace=TRUE),
       Y1=sample(c(NA,LETTERS[1:4]),10,replace=TRUE),
       Y1=sample(c(NA,c(9,7,5)),10,replace=TRUE),
       Y1=sample(c(NA,c("Y","N")),10,replace=TRUE))
X = data.frame(X=sample(c("Y","N"),10,replace=TRUE),
       X=sample(1:10,10,replace=TRUE),
       X=sample(LETTERS[23:26],10,replace=TRUE),
       X=sample(c(11,22,33),10,replace=TRUE),
       X=sample(0:1,10,replace=TRUE),
       X=sample(c("M","F"),10,replace=TRUE))
```

Imputation in R

```
#Need a bigger dataset for the mathematical approaches
set.seed(111)
Y2 = data.frame(Y1=sample(c(NA,1:5),100,replace=TRUE),
  Y1=sample(c(NA,LETTERS[1:4]),100,replace=TRUE),
  Y1=sample(c(NA,c(9,7,5)),100,replace=TRUE),
  Y1=sample(c(NA,c("Y","N")),100,replace=TRUE,prob=c(0.2,0.4,0.4)))
X2 = data.frame(X=sample(c("Y","N"),100,replace=TRUE),
  X=sample(1:10,100,replace=TRUE),
  X=sample(LETTERS[23:26],100,replace=TRUE),
  X=sample(c(11,22,33),100,replace=TRUE),
  X=sample(0:1,100,replace=TRUE),
  X=sample(c("M","F"),100,replace=TRUE))

YFull = rbind(Y,Y2)
XFull = rbind(X,X2)
```

Imputation in R

```
#mean imputation
impute.mean = function(var){
  ind = which(is.na(var))
  ind2 = which(!is.na(var))
  if(class(var)%in%c("integer","numeric")){
    m=mean(var[ind2])
    var[ind]=m
  }
  else if(class(var)=='factor'){
    tab = sort(table(var))
    tab = tab[which(tab==max(tab))]
    var[ind] = sample(names(tab),length(ind),replace=TRUE)
  }
  var
}
Y.meanImpute = do.call(cbind.data.frame,lapply(Y,impute.mean))
```

Imputation in R


```
#random imputation
impute.random = function(var){
  ind = which(is.na(var))
  ind2 = which(!is.na(var))
  replace = sample(var[ind2],length(ind),replace=TRUE)
  var[ind] = replace
  var
}

Y.randImpute = do.call(cbind.data.frame,lapply(Y,impute.random))
```

Imputation in R

```
#regression imputation
Y.regress = Y

mod01 = lm(YFull[,1]~.,data=XFull)
pred = predict(mod01,newdat=X)
ind = which(is.na(Y[,1]))
Y.regress[ind,1] = pred[ind]

mod04 = glm(YFull[,4]~.,family='binomial',data=XFull)
pred = predict(mod04,newdat=X)
pred = c("Y","N")[(exp(pred)/(1+exp(pred))>0.5)+1]
ind = which(is.na(Y[,4]))
Y.regress[ind,4] = pred[ind]
```

Imputation in R

```
#hot deck imputation
library(hot.deck)
temp = hot.deck(cbind(YFull,XFull))
Y.hotdeck = temp[['data']][[1]][1:10,1:4]
print(xtable(Y.hotdeck,digits=1),include.rownames=FALSE)
```

Imputation in STATA

- It appears that the the multiple imputation command `mi` is used for all imputation in STATA now
  - There is a command called `impute` but it is depreciated
- mi impute is the command to perform imputation on a single variable
- **need to be careful:** Once you've started using `mi` there's a risk of performing your analysis very incorrectly

In STATA


```
//testing imputation using STATA default code
drop _all
use http://www.stata-press.com/data/r13/mheart1s0

regress bmi attack smokes age female hsgrad

mi impute regress bmi attack smokes age female hsgrad,
    add(1)

regress bmi attack smokes age female hsgrad

mi impute regress bmi attack smokes age female hsgrad,
    add(20)

regress bmi attack smokes age female hsgrad
mi estimate: regress bmi attack smokes age female
   hsgrad
```

Second STATA example

```
//another imputation example from: https://stats.idre.
    ucla.edu/stata/seminars/mi_in_stata_pt1_new/
drop _all
use https://stats.idre.ucla.edu/wp-content/uploads
    /2017/05/hsb2_mar.dta, clear

sum

regress ses i.female
```

Second STATA example

```
//required before we start imputing
mi set mlong

mi misstable summarize female write read math prog

//need to register the variable to impute
mi register imputed female

//simple imputation of female
mi impute logit female race schtyp socst, add(1)

regress ses i.female

// now to perform multiple imputation
mi impute logit female race schtyp socst, add(30)
regress ses i.female
mi estimate: regres ses i.female
```

Select Output

```
. sum

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          id |        200       100.5    57.87918         1        200
      female |        182   .5549451    .4983428         0          1
        race |        200        3.43    1.039472         1          4
         ses |        200       2.055    .7242914         1          3
      schtyp |        200        1.16     .367526         1          2
-------------+---------------------------------------------------------
        prog |        182   2.027473    .6927511         1          3
        read |        191   52.28796    10.21072        28         76
       write |        183   52.95082    9.257773        31         67
        math |        185    52.8973    9.360837        33         75
     science |        184   51.30978    9.817833        26         74
-------------+---------------------------------------------------------
       socst |        200     52.405    10.73579        26         71
```

Select Output

```
. regress ses i.female
      Source |       SS           df       MS      Number of obs   =       182
-------------+----------------------------------   F(1, 180)       =      2.37
       Model |  1.19339034         1  1.19339034   Prob > F        =    0.1253
    Residual |  90.5703459       180  .503168588   R-squared       =    0.0130
-------------+----------------------------------   Adj R-squared   =    0.0075
       Total |  91.7637363       181  .506981968   Root MSE        =    .70934

------------------------------------------------------------------------------
         ses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
     female  |  -.1629385   .1058009    -1.54   0.125    -.3717081     .045831
       _cons |    2.17284    .078816    27.57   0.000     2.017317    2.328362
------------------------------------------------------------------------------

. regress ses i.female race schtyp socst
      Source |       SS           df       MS      Number of obs   =       182
-------------+----------------------------------   F(4, 177)       =      8.88
       Model |  15.3422745         4  3.83556862   Prob > F        =    0.0000
    Residual |  76.4214618       177  .431759671   R-squared       =    0.1672
-------------+----------------------------------   Adj R-squared   =    0.1484
       Total |  91.7637363       181  .506981968   Root MSE        =    .65708

------------------------------------------------------------------------------
         ses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
     female  |  -.2060478   .0984047    -2.09   0.038    -.4002452   -.0118504
        race |   .0818726   .0492276     1.66   0.098    -.0152759    .1790211
      schtyp |   .1890819   .1341215     1.41   0.160    -.0756012    .4537649
       socst |   .0225477   .0048352     4.66   0.000     .0130057    .0320898
       _cons |   .5036557   .3069852     1.64   0.103    -.1021665    1.109478
```

Select Output

```
. mi misstable summarize female write read math prog
                                                              Obs<.
                                                +------------------------------
                |                               | Unique
     Variable   |     Obs=.     Obs>.     Obs<. | values        Min          Max
  --------------+-----------------------------------+------------------------------
         female |        18                   182 |      2          0            1
          write |        17                   183 |     29         31           67
           read |         9                   191 |     30         28           76
           math |        15                   185 |     39         33           75
           prog |        18                   182 |      3          1            3
  ------------------------------------------------------------------------------

. mi impute logit female race schtyp socst, add(1)

Univariate imputation                       Imputations =        1
Logistic regression                               added =        1
Imputed: m=1                                    updated =        0

------------------------------------------------------------------
                    |                Observations per m
                    |-----------------------------------------
          Variable  |   Complete   Incomplete    Imputed |     Total
--------------------+-----------------------------------+----------
              female |        182           18         18 |       200
------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

Select Output

```
. regress ses i.female

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(1, 198)       =      3.19
       Model |  1.65459596         1  1.65459596   Prob > F        =    0.0757
    Residual |  102.740404       198  .518890929   R-squared       =    0.0158
-------------+----------------------------------   Adj R-squared   =    0.0109
       Total |     104.395       199   .52459799   Root MSE        =    .72034

------------------------------------------------------------------------------
         ses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
     female  |  -.1828283   .1023848    -1.79   0.076    -.3847329    .0190763
       _cons |   2.155556   .0759306    28.39   0.000     2.005819    2.305292
------------------------------------------------------------------------------

. regress ses i.female race schtyp socst

      Source |       SS           df       MS      Number of obs   =       200
-------------+----------------------------------   F(4, 195)       =      9.63
       Model |  17.2269908         4  4.3067477   Prob > F        =    0.0000
    Residual |  87.1680092       195  .447015432   R-squared       =    0.1650
-------------+----------------------------------   Adj R-squared   =    0.1479
       Total |     104.395       199   .52459799   Root MSE        =    .66859

------------------------------------------------------------------------------
         ses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
     female  |  -.2450124    .095694    -2.56   0.011    -.4337404   -.0562844
        race |   .0912012   .0466452     1.96   0.052    -.0007926    .183195
      schtyp |   .1844124   .1301195     1.42   0.158    -.0722097   .4410346
```

Select Output

```
. mi impute logit female race schtyp socst, add(30)

Univariate imputation                          Imputations =       31
Logistic regression                                  added =       30
Imputed: m=2 through m=31                          updated =        0

------------------------------------------------------------------------
                     |                Observations per m
                     |-----------------------------------------------
            Variable |   Complete   Incomplete   Imputed |      Total
---------------------+-----------------------------------+-----------
              female |        182           18        18 |        200
------------------------------------------------------------------------
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)
```

Select Output

```
. regress ses i.female

      Source |       SS           df       MS      Number of obs   =       740
-------------+----------------------------------   F(1, 738)       =      0.71
       Model |  .434468729         1  .434468729   Prob > F        =    0.3981
    Residual |  448.510126       738  .607737298   R-squared       =    0.0010
-------------+----------------------------------   Adj R-squared   =   -0.0004
       Total |  448.944595       739  .607502834   Root MSE        =    .77958

------------------------------------------------------------------------------
         ses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
     female  |  -.0486321   .0575177    -0.85   0.398      -.16155    .0642858
       _cons |   1.879056   .0423407    44.38   0.000     1.795933    1.962179
------------------------------------------------------------------------------
```

Select Output

```
. mi estimate: regres ses i.female

Multiple-imputation estimates                     Imputations       =         31
Linear regression                                 Number of obs     =        200
                                                  Average RVI       =     0.0728
                                                  Largest FMI       =     0.1281
                                                  Complete DF       =        198
DF adjustment:   Small sample                     DF:     min       =     157.12
                                                          avg       =     165.71
                                                          max       =     174.31
Model F test:       Equal FMI                     F(   1,  157.1)   =       1.82
Within VCE type:          OLS                     Prob > F          =     0.1797

------------------------------------------------------------------------------
         ses |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |
     female  |  -.1480762   .1098837    -1.35   0.180     -.365116    .0689636
       _cons |    2.13705   .0795989    26.85   0.000     1.979948    2.294151
------------------------------------------------------------------------------
```

Comparing Results

	n	β	SE	t	p-value	[95% CI]	
Complete	182	-.16293	.1058009	-1.54	0.125	-.371708	.045831
Imputed	200	-.18282	.1023848	-1.79	0.076	-.384732	.019076
MI	*	-.14807	.1098837	-1.35	0.180	-.365116	.068963
WRONG	740	-.04863	.0575177	-0.85	0.398	-.161550	.064285

- The three correct estimates are relatively similar
- We can see in the MI example the proper effect of imputation on standard error
  - I didn't explain how `mi estimate` does this, but the help file covers it if you're interested `http://www.stata.com/manuals13/mimiestimate.pdf`