Using R for Analytic Graphs: Learn How Data Visualization Can Improve Interpretation in Social Work Research

Saturday, November 01, 2014

Why Use R?

- Free
- Open Source
- Easy Collaboration
- Replicable Research

Why Wouldn't You Use R?

Steep(er) learning curve compared to, say, Excel or SPSS. This matters a lot if

- You run statistics rarely.
- You want a point and click interface.

Where Can you Get R?

- CRAN
- Our Thumb Drives

Where Are We Going Today?

- Graphing Descriptive Statistics
- Graphing Model Results

Graphing Descriptive Statistics

What we see in the journal article. . .

Placement status	Aboriginal (%)	Caucasian (%)
Child welfare placement	9.90	4.6
Informal placement	11.20	3.4
Placement considered	3.90	2.4
No placement required	75.10	89.6
N	831.00	3,563

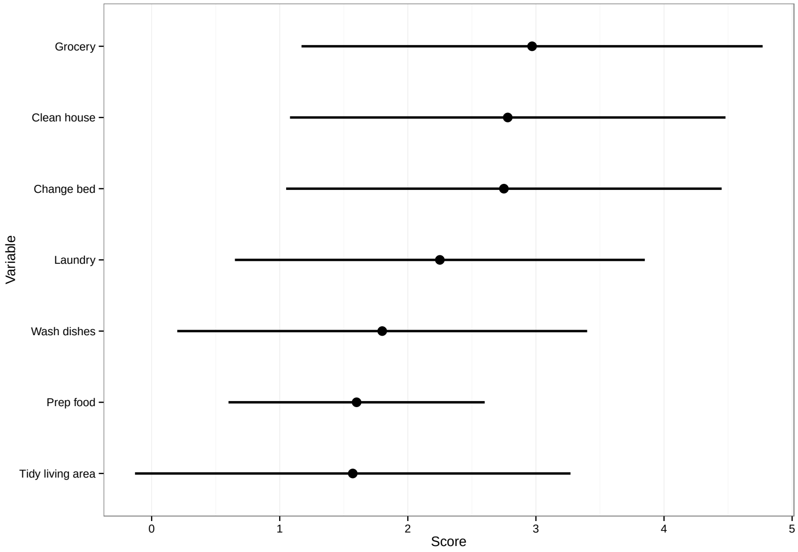
Graphing Descriptive Statistics

What we see in the journal article...

Subscale and Range	Abbreviated item and Content	M(+-SD)	Item-total correlation
Household Activities	Prepare food	1.60(1.00)	0.72
	Grocery shop	2.97(1.80)	0.78
	Clean house	2.78(1.70)	0.82
	Laundry	2.25(1.60)	0.83
	Change bed linen	2.75(1.70)	0.83
	Wash dishes	1.80(1.30)	0.72
	Tidy living area	1.57(1.20)	0.74

Graphing Descriptive Statistics

What we could see in the journal article...



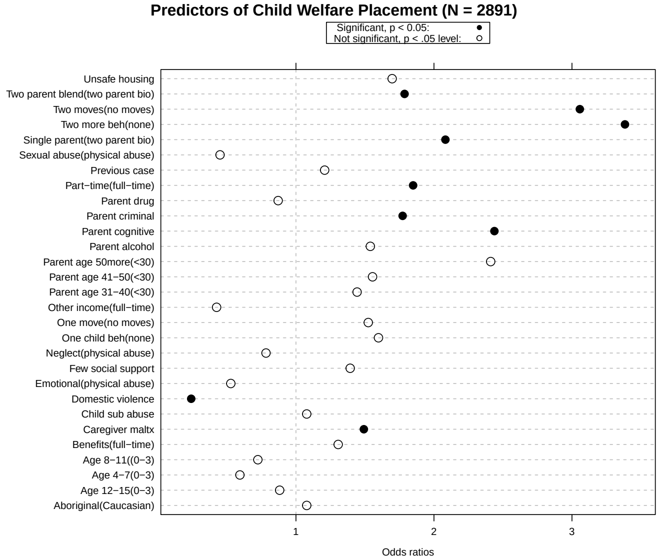
Graphing Model Results

What we see in the journal article...

Variable	Coefficient	OR	pvalue
Aboriginal(Caucasian)	0.07	1.08	0.74
Part-time(full-time)	0.61	1.85	0.04
Benefits(full-time)	0.27	1.31	0.30
Other income(full-time)	-0.86	0.42	0.14
Unsafe housing	0.53	1.70	0.07
One move(no moves)	0.42	1.52	0.06
Two moves(no moves)	1.12	3.06	0.00
Two parent blend(two parent bio)	0.58	1.79	0.05
Single parent(two parent bio)	0.73	2.08	0.00
Previous case	0.19	1.21	0.34
Sexual abuse(physical abuse)	-0.80	0.45	0.05
Neglect(physical abuse)	-0.24	0.78	0.26
Emotional(physical abuse)	-0.64	0.53	0.06
Domestic violence	-1.42	0.24	0.00
Child sub abuse	0.07	1.08	0.84
One child beh(none)	0.47	1.60	0.70
Two more beh(none)	1.22	3.38	0.00
Age 4-7(0-3)	-0.52	0.59	0.06
Age 8-11((0-3)	-0.32	0.72	0.28
Age 12-15(0-3)	-0.12	0.88	0.71
Parent age 31-40(<30)	0.37	1.44	0.12
Parent age 41-50(<30)	0.44	1.55	0.21
Parent age 50more(<30)	0.88	2.41	0.32
Parent drug	-0.14	0.87	0.59
Parent criminal	0.57	1.77	0.02
Parent cognitive	0.89	2.44	0.00
Few social support	0.33	1.39	0.07
Caregiver maltx	0.40	1.49	0.04
Parent alcohol	0.43	1.54	0.05

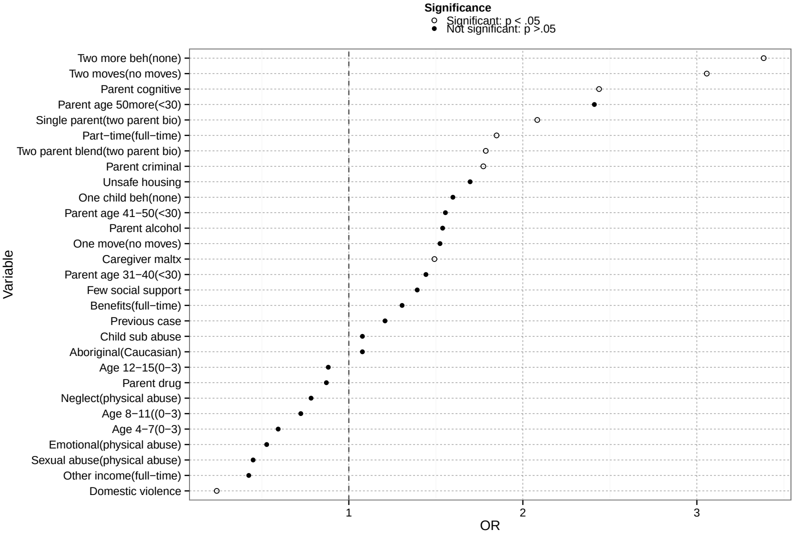
Graphing Model Results

What we could see in the journal article. . .



Graphing Model Results

What we could see in the journal article. . .



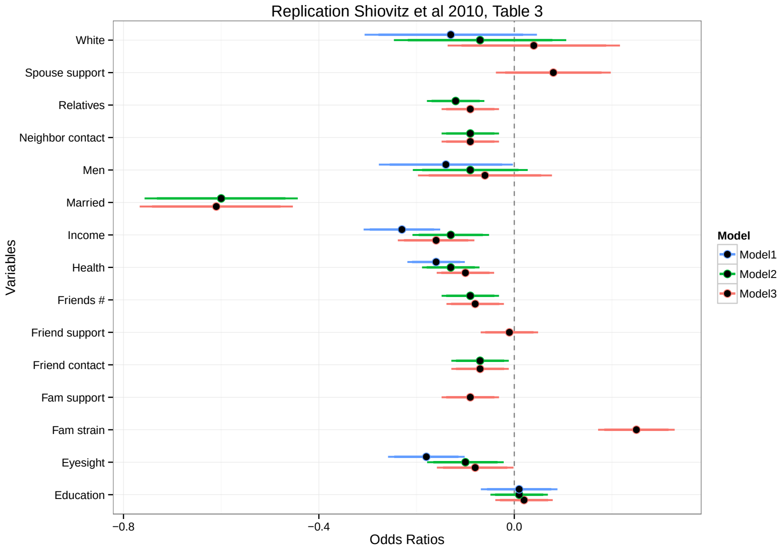
Graphing Model Results

What we see in the journal article...

Variable	Coef	SE	p	Coef	SE	p	Coef	SE	p
Education	0.01	0.04	0.2	0.01	0.03	0.2	0.02	0.03	0.20
Eyesight	-0.18	0.04	0.001	-0.1	0.04	0.01	-0.08	0.04	0.05
Health	-0.16	0.03	0.001	-0.13	0.03	0.001	-0.1	0.03	0.01
Income	-0.23	0.04	0.001	-0.13	0.04	0.01	-0.16	0.04	0.01
Men	-0.14	0.07	0.2	-0.09	0.06	0.2	-0.06	0.07	0.20
White	-0.13	0.09	0.2	-0.07	0.09	0.2	0.04	0.09	0.20
Friend contact				-0.07	0.03	0.1	-0.07	0.03	0.10
Neighbor contact				-0.09	0.03	0.001	-0.09	0.03	0.00
Married				-0.6	0.08	0.001	-0.61	0.08	0.00
Friends #				-0.09	0.03	0.05	-0.08	0.03	0.10
Relatives				-0.12	0.03	0.001	-0.09	0.03	0.01
Fam strain							0.25	0.04	0.00
Fam support							-0.09	0.03	0.01
Friend support							-0.01	0.03	0.20
Spouse support							0.08	0.06	0.20

Graphing Model Results

What we could see in the journal article...



Graphing Model Results

What if we had more information than what was available in peer-reviewed journals?

Consider this basic algorithm

1. Choose a counterfactual x_c.
2. Estimate a model to get a vector of parameters $\hat{\beta}$ and the associated variance-covariance matrix, $\hat{\boldsymbol{V}}$.
3. Draw several $\tilde{\beta}$ from $\mathcal{N}(\hat{\beta}, \hat{\boldsymbol{V}})$, where $\mathcal{N}$ is a mulivariate normal distribution.
4. Calculate expected outcomes based on model parameters for all of your draws from $\mathcal{N}$.
5. Calculate summary statistics for each level of x_c.

This approach will work for most of the models that social welfare researchers tend to encounter.

A Practical Example - Background

Research Question

How does a child's probability of exiting the foster care system vary by child characteristics?

Multiple Permanency Outcomes

Requires that we estimate a mulinomial logistic regression model.

Data in Question

- 500 children entering out-of-home care in late 2007.
- Children's parent's were surveyed once in 2007. The survey results were then linked to administrative data which faciliated a longitudinal follow-up.
- Data have been jittered and randomly sampled from a larger set of data to mask the identity of subjects. The data used here do not reflect the data of individual subjects.

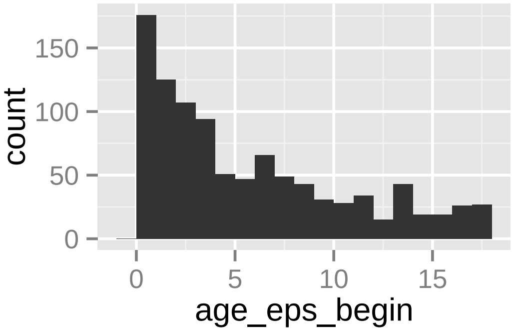
A practical example - Choose a counterfactual x_c.

Load the data

```
dat <- read.csv("dat.csv")
```

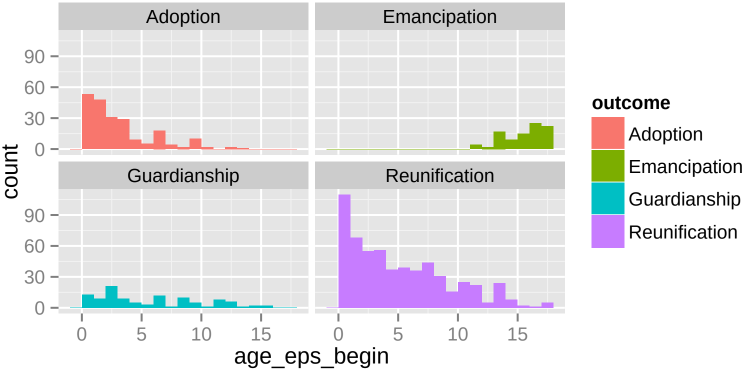
A practical example - Choose a counterfactual x_c.

```r
#looking at age of child at episode begin
require(ggplot2)
ggplot(dat, aes(x=age_eps_begin)) +
  geom_histogram(binwidth = 1)
```



A practical example - Choose a counterfactual x_c.

```
#looking at age of child at episode begin by outcome
ggplot(dat, aes(x=age_eps_begin, fill=outcome)) +
  geom_histogram(binwidth = 1) +
  facet_wrap(~ outcome)
```



A practical example - Estimate a model.

Need to estimate a statistical model to get

1. A vector of parameters $\hat{\beta}$, and
2. The associated variance-covariance matrix, $\hat{\boldsymbol{V}}$.

A practical example - Estimate a model.

Prep the data

```r
# easy to load external packages
# install.packages("nnet") # install once
require(nnet)                # load every time

# relevel our outcome variable
dat$outcome_rl <- relevel(dat$outcome
                          , ref = "Emancipation")

# recode to numeric
dat$outcome_rl <- as.numeric(dat$outcome_rl)
```

A practical example - Estimate a model.

Run the model

```
# run the multinomial model
model <- multinom(outcome_rl ~ age_eps_begin +
                   eps_rank
                  ,data = dat
                  ,Hess = TRUE)
```

```
## # weights:  16 (9 variable)
## initial  value 1386.294361
## iter  10 value 931.103300
## iter  20 value 860.375750
## final  value 860.374425
## converged
```

A practical example - Estimate a model.

Display of summary the model

```
model
```

```
## Call:
## multinom(formula = outcome_rl ~ age_eps_begin + eps_rank
##     Hess = TRUE)
##
## Coefficients:
##   (Intercept) age_eps_begin    eps_rank
## 2   11.457365    -1.0280750 -0.10995325
## 3    9.797665    -0.8393067  0.05195097
## 4   11.597181    -0.8691345  0.07149574
##
## Residual Deviance: 1720.749
## AIC: 1738.749
```

A practical example - Estimate a model.

Extract a vector of parameters $\hat{\beta}$

```
#run the multinomial model
pe <- model$wts[c(6,7,8,10,11,12,14,15,16)]
pe[1:3]
```

```
## [1] 11.4573653 -1.0280750 -0.1099532
```

```
pe[4:6]
```

```
## [1]  9.79766546 -0.83930667  0.05195097
```

```
pe[7:9]
```

```
## [1] 11.59718150 -0.86913446  0.07149574
```

A practical example - Estimate a model.

Extract the associated variance-covariance matrix, $\hat{\boldsymbol{V}}$

```
#run the multinomial model
vc <- solve(model$Hess)
```

A practical example - Draw several $\tilde{\beta}$ from $\mathcal{N}(\hat{\beta},\ \hat{\boldsymbol{V}})$.

```
#load a package which contains a multivariate normal
#sampling function
require(MASS)
#assign a variable for the number of simulations
sims <- 10000
#draw the indicates number of beta simulates
#using our extracted model data
simbetas <- mvrnorm(sims,pe,vc)
```

A practical example - Last two steps...

- Calculate expected values for all of your draws from $\mathcal{N}$, and
- Calculate summary statistics for each level of x_c.
- Specific calculations are beyond the scope of this presentation
- But the `simcf` package from Chris Adolph (political scientist at the University of Washington) will do them for us!

A practical example - Last two steps

Get data read for `simcf`

- Re-arrange simulates to array format

```
simb <- array(NA, dim = c(sims,3,3))
simb[,,1] <- simbetas[,1:3]
simb[,,2] <- simbetas[,4:6]
simb[,,3] <- simbetas[,7:9]
```

- Specify range of counterfactual values

```
agerange <- seq(0,17,by=0.1)
```

A practical example - Last two steps

Get data read for simcf

- Load simcf and use the cfFactorial() function to set specific values for simulation.

```
require(simcf)
xhyp <- cfFactorial(age = agerange
                    ,ep_rank = mean(dat$eps_rank))
```

- Run the simulation (this is where the last two steps are really performed).

```
test_sims <- mlogitsimev(xhyp,simb,ci=0.95)
```

Get the data ready to graph

```
y <- as.vector(test_sims$pe[,1:4])

x <- rep(1:length(agerange), 4)

lower <- as.vector(test_sims$lower[,1:4,])

upper <- as.vector(test_sims$upper[,1:4,])

Outcome <- c(rep("Adoption", length(agerange))
                ,rep("Guardianship"
                     ,length(agerange))
                ,rep("Reunification"
                     ,length(agerange))
                ,rep("Emancipation"
                     ,length(agerange)))
```

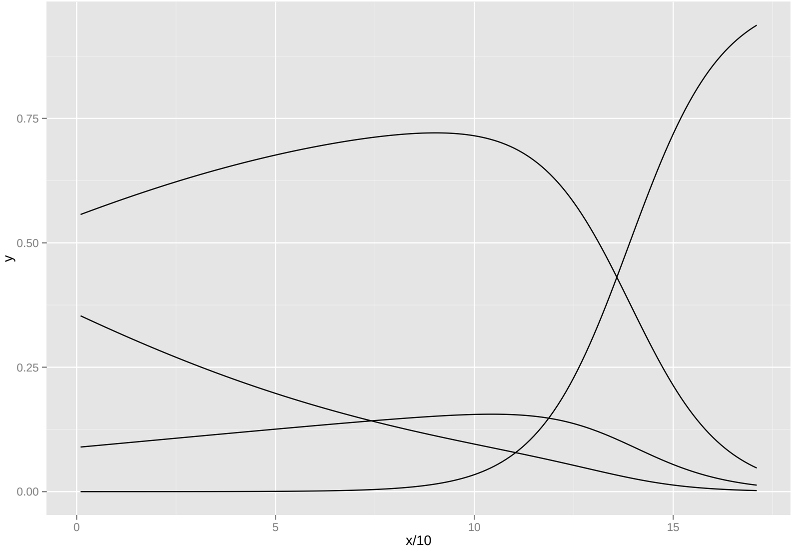
Get the data ready to graph

```
dat_sim_plot <- data.frame(y,x,lower,upper,Outcome)
```

Graph the data!

```
p1 <- ggplot(dat_sim_plot
        ,aes(x=x/10, y=y, group=Outcome)) +
         geom_line()
```

Graph the data!



Make it Pretty!

```
p2 <- ggplot(dat_sim_plot
       ,aes(x=x/10, y=y, group=Outcome)) +
        geom_line(size=1, alpha=.5) +
        geom_ribbon(aes(ymin=lower
                         ,ymax=upper
                         ,fill=Outcome), alpha=.5) +
        ylab("Pr(Outcome|Age,Prior Episodes)") +
        xlab("Age at Entry into Foster Care") +
        theme_bw()
```

Make it Pretty