36-617: Applied Linear Models

Two Vignettes:

Meta-Analysis and Extending Multilevel GLM's

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Announcements

- **HW10** Due Weds 1159
- **Carnegie Mellon Faculty Course Evaluations (FCE's)** Go to https://cmu.smartevals.com/, find our class, and rate for CMU
- **Last two classes**
  - Monday: Two vignettes: Sensitivity, and Correlated Etas
  - Today: Two vignettes: Meta Analysis and Extending Multilevel glm's
- **Final Papers** Due Friday 1159pm

Outline

- Vignette 1: Meta Analysis
  - Example: Diet and Exercise vs Cholesterol
- Vignette 2: Multilevel Poisson regression and extensions
  - Example: the roaches data

Meta-Analysis

- Kelly et al. (2011). Efficacy of aerobic exercise and a prudent diet for improving selected lipids and lipoproteins in adults: a meta-analysis of randomized controlled Trials. *BMC Medicine 9:74* http://www.biomedcentral.com/1741-7015/9/74
  - Question: Can exercise and diet lower LDL cholesterol and raise HDL cholesterol?
  - Reviewed 1,401 papers, found 6 studies that were RCT's with pt. estimates and CI's for the effect of exercise and diet on cholesterol.

Meta-analysis = review of papers on a single phenomenon. Can be qualitative, quantitative, or both.

RCT = "Randomized Controlled Trial"

We'll focus on HDL-C as Example…

Study	Subjects	Assignment[1]	Subgroup	Effect[2]	CI
Hellenuis et al. (1993)	78 men, 35–60 y.o.	39 Tx, 39 Ctrl	All	−0.4	(−0.9, 0.2)
McAuley et al. (2002)	52 men & women, 30–68 y.o.	29 Tx, 23 Ctrl	All	−5.0	(−5.5, −4.6)
Miller et al. (2002)	43 hypertensive, overweight, 22–70 y.o.	20 Tx, 23 Ctrl	All	−5.0	(−8.0, −2.0)
Neiman et al. (2002)	44 sedentary obese women, 25–75 y.o.	22 Tx, 22 Ctrl	All	−2.3	(−2.8, −1.9)
Stefanick et al. (1998)	182 men & women, 30–64 y.o.	91 Tx, 91 Ctrl	Men	0.6	(−1.4, 2.6)
Stefanick et al. (1998)	– –	– –	Women	−2.1	(−4.7, 0.5)
Wood et al. (1991)	160 sendentary overweight men & women, 25–49 y.o.	81 Tx, 79 Ctrl	Men	7.3	(6.9, 7.8)
Wood et al. (1991)	– –	– –	Women	2.7	(2.1, 3.3)

[1]Randomly assigned to either Tx (diet + exercise) or Ctrl (no intervention).
[2]Effect is $\overline{\text{HDL-C}}_{Tx} - \overline{\text{HDL-C}}_{Ctrl}$. Negative values favor Tx (treatment).
Source: Kelly et al. (2011), Table 1 & Figure 4.

- Want to combine these to get an overall estimate of the effect of diet & exercise on HDL-C
- We can use an MLM to do a "random-effects meta-analysis"…

Kelly et al. (2011) also address LDL-C, but HDL-C presents more interesting data for analysis.

The *ideal* MLM…

- In a simple randomized controlled study (RCT), we could fit the linear model

$$\text{HDL}-\text{C}_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

  where $X_i = 1$ for Tx, and $X_i = 0$ for Ctrl.
  - We know $\hat{\beta}_1$ estimates the treatment effect.
- So if we had access to the individual participants' HDL levels, we could use an MLM like

$$\text{HDL}-\text{C}_i = \alpha_{0j[i]} + \alpha_{1j[i]} X_i + \epsilon_i$$
$$\alpha_{0j[i]} = \beta_0 + \eta_{0j}$$
$$\alpha_{1j[i]} = \beta_1 + \eta_{1j}$$

  where $j[i]$ is the study that participant i was in.
  - $\hat{\beta}_1$ would be the overall Tx effect
  - $\hat{\alpha}_{1j}$ would be the Tx effect for study j.
- *But we don't have individual participants' HDL levels…*

What do we *actually* have to work with…

- Let $\theta_j = \mathrm{E}[\mathrm{HDL{-}C}|\mathrm{Tx}] - \mathrm{E}[\mathrm{HDL{-}C}|\mathrm{Ctrl}]$
  - Like α_{1j} in the "full" MLM…
- The estimates in the studies give us $\hat{\theta}_j$ and $SE(\hat{\theta}_j)$:
  - $\hat{\theta}_j$ = estimate in the j[th] study = $\overline{\mathrm{HDL{-}C}}_{Tx} - \overline{\mathrm{HDL{-}C}}_{Ctrl}$
  - $SE(\hat{\theta}_j) \approx$ (width of CI)/4
    - Why: sample size ⇒ CLT, and 1.96 ≈ 2.

Study	Subgroup	Effect ($\hat{\theta}_j$)	CI	$SE(\hat{\theta}_j)$
Hellenuis et al. (1993)	All	−0.4	(−0.9, 0.2)	0.275
McAuley et al. (2002)	All	−5.0	(−5.5, −4.6)	0.225
Miller et al. (2002)	All	−5.0	(−8.0, −2.0)	1.5
Neiman et al. (2002)	All	−2.3	(−2.8, −1.9)	0.225
Stefanick et al. (1998)	Men	0.6	(−1.4, 2.6)	1.0
Stefanick et al. (1998)	Women	−2.1	(−4.7, 0.5)	1.3
Wood et al. (1991)	Men	7.3	(6.9, 7.8)	0.225
Wood et al. (1991)	Women	2.7	(2.1, 3.3)	0.3

There can be better ways to estimate $SE(\hat{\theta}_j)$; see Follman, D., Elliot, P., Suh, I. and Cutler, J. (1992). Variance imputation for overviews of clinical trials with continuous response. *Journal of Clinical Epidemiology, 45,* 769--773.

The MLM we *can* build…

- This suggests the simpler MLM

lmer() can't do this!)

$$\hat{\theta}_j \sim N\left(\theta_j, \tau_j^2\right)$$
$$\theta_j \sim N(\beta_1, \sigma^2)$$

where $\hat{\theta}_j$ and $\tau_j \approx SE(\hat{\theta}_j)$ can be plugged in from the table on the previous slide.

```
data{
  int J;
  real theta_hat[J];
  real tau[J];
}

parameters {
  real theta[J];
  real beta1;
  real<lower=0, upper=50> sigma;
}
```

```
model {
for (j in 1:J) {
  theta_hat[j] ~
normal(theta[j],tau[j]);
  }

for (j in 1:J) {
  theta[j] ~ normal(beta1,sigma);
  }

beta1 ~ normal(0,1e+6);
sigma ~ uniform(0,50);
}
```


This stan model is in the file RE-meta-analysis.stan

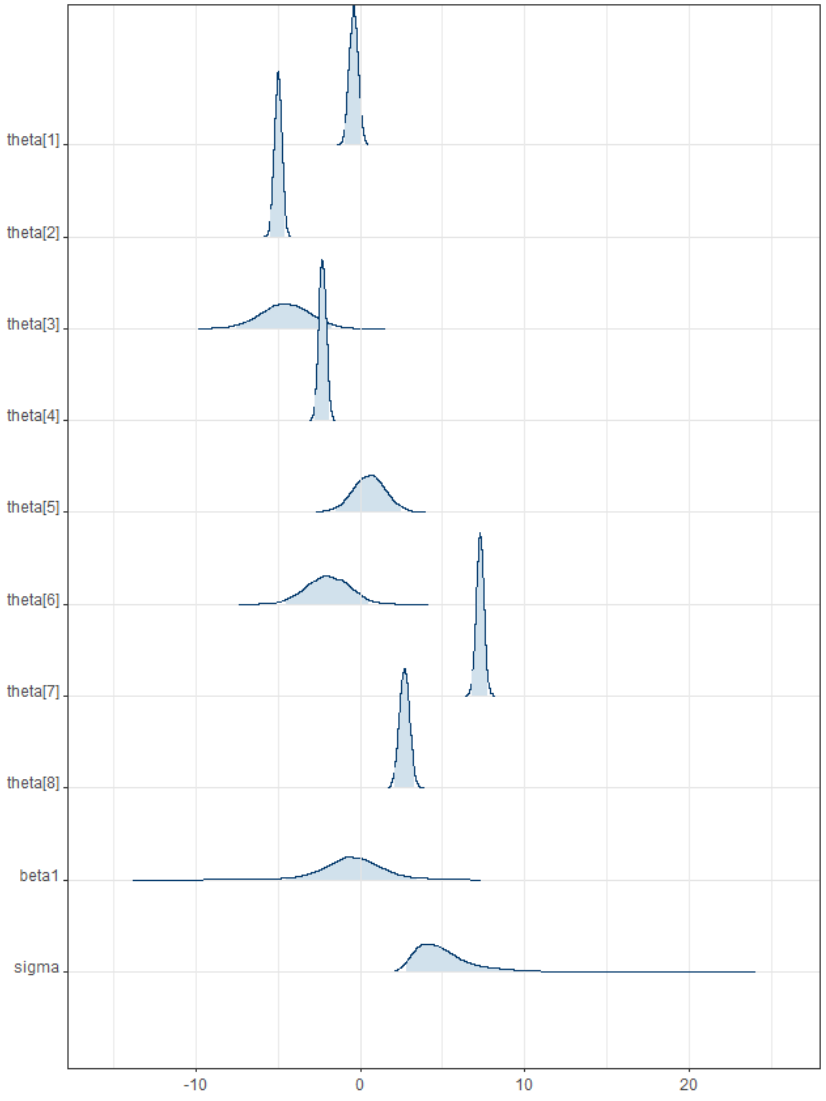
This is called a ***random effects meta-analysis model*** because it treats the individual study effects θ_j as random effects.

Fitting the model…

```
## code to set up data frame raw_data
## is in RE-meta-analysis.r
metaanalysis_data <- c(list(J=J),
                       as.list(raw_data))
meta_model <- stan(
  file="RE-meta-analysis.stan",
  data=metaanalysis_data,chains=0)
meta_result <-stan(
  fit=meta_model,
  data=metaanalysis_data)
meta_result
```

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
theta[1]	-0.40	0.00	0.27	-0.91	-0.59	-0.40	-0.22	0.12	7141	1
theta[2]	-4.99	0.00	0.22	-5.42	-5.14	-4.99	-4.84	-4.56	7873	1
theta[3]	-4.54	0.02	1.48	-7.48	-5.53	-4.54	-3.54	-1.72	7575	1
theta[4]	-2.29	0.00	0.22	-2.72	-2.45	-2.30	-2.14	-1.86	7129	1
theta[5]	0.54	0.01	1.00	-1.48	-0.13	0.56	1.21	2.52	7468	1
theta[6]	-1.98	0.01	1.31	-4.50	-2.87	-1.97	-1.07	0.52	9703	1
theta[7]	7.28	0.00	0.23	6.83	7.12	7.28	7.43	7.73	8156	1
theta[8]	2.69	0.00	0.30	2.10	2.48	2.69	2.89	3.27	8028	1
beta1	-0.48	0.03	1.92	-4.53	-1.58	-0.48	0.68	3.43	3466	1
sigma	5.07	0.04	1.86	2.79	3.82	4.69	5.78	9.52	2553	1
lp__	-18.66	0.06	2.39	-24.26	-19.97	-18.25	-16.91	-15.11	1514	1

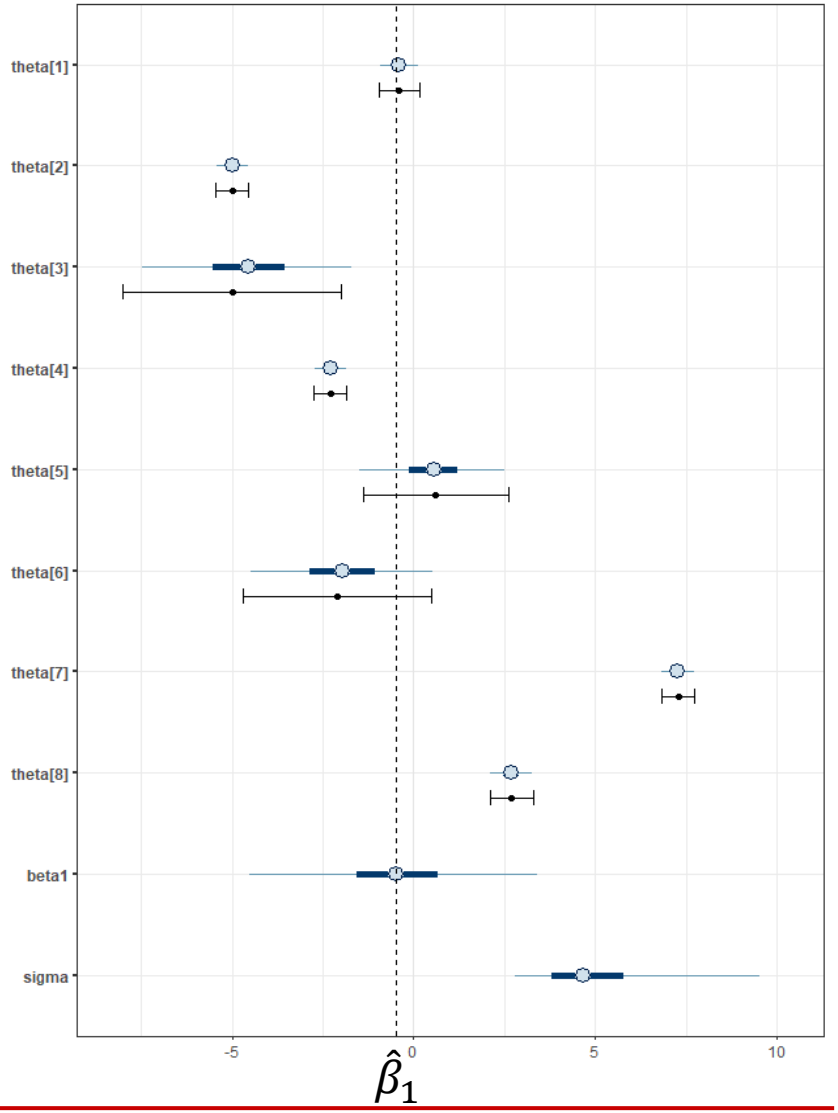
```
mcmc_areas_ridges(meta_result,
    pars=c(paste0("theta[",1:J,"]"),
           "beta1","sigma"),
           prob_outer=0.95)
```



This R code is in the file RE-meta-analysis.r

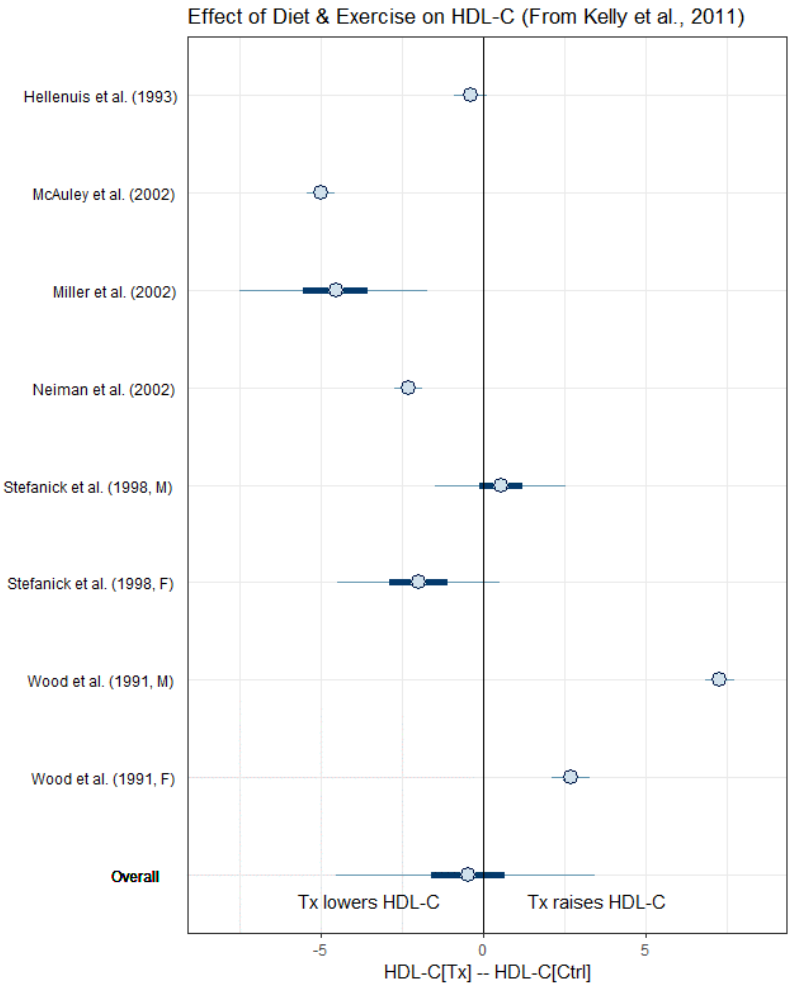
Comparing Raw intervals to Bayes…

- Raw data vs. Bayes
  - Bayes: Blue intervals
  - Raw data: Black intervals
- Shrinkage toward $\hat{\beta}_1$ for studies with larger SE's
  - Especially for θ_3 and θ_6
- Narrower intervals for Bayes vs Raw Data
  - Borrowing information from other studies makes each study estimate more precise



Reporting the results…

- This sort of plot is called a "Forest Plot" in meta-analysis
  - Replaced θ_1, θ_2, etc. with study names
  - Replaced β_1 with "Overall"
- We see that the CI for "Overall" covers 0, so we can't conclude that "diet & exercise" has a significant effect on HDL-C
- The McAuley & Wood studies seem to be influential. Possible next steps:
  - Re-analyze without each of the 8 studies; get same results? (LOOCV!)
  - Investigate why results are different

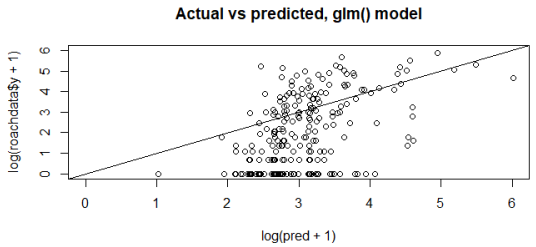


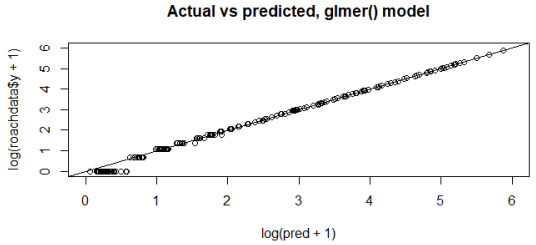
Extending Multi-Level GLM's: The Roach Data

```
str(roachdata)
#
# 'data.frame':   262 obs. of  6 variables:
#  $ X         : int  1 2 3 ...  [observation number]
#  $ y         : int  153 127 ... [# of roaches post-expmt]
#  $ roach1    : num  308 ...     [# of roaches pre-expmt]
#  $ treatment: int  1 1 1 1 ... [pest mgmt tx in this apt bldg?]
#  $ senior    : int  0 0 0 0 ... [apts restricted to sr citzns?]
#  $ exposure2: num  0.8 0.6 ... [avg # of trap-days per apt]
```

Review of our glmer() analyses…

```
glm.1 <- glm (y ~ roach1 + treatment +
  senior, offset=log(exposure2),
  family=poisson, data=roachdata)

glmer.2 <- glmer(y ~ roach1 + treatment +
  senior + (1|X), offset=log(exposure2),
  family=poisson,
  data=roachdata)

par(mfrow=c(2,1))

pred <- predict(glm.1,type="response")
plot(log(pred+1),log(roachdata$y+1),xlim=
c(0,6),ylim=c(0,6))
title(main="Actual vs predicted, glm()
model")
abline(0,1)

pred <- predict(glmer.2,type="response")
plot(log(pred+1),log(roachdata$y+1),xlim=
c(0,6),ylim=c(0,6))
title(main="Actual vs predicted, glmer()
model")
abline(0,1)
```






```
display(glmer.2)
##             coef.est coef.se
## (Intercept)  1.08     0.27
## roach1        0.02     0.00
## treatment    -0.82     0.31
## senior       -1.04     0.35
##
## Error terms:
##  Groups   Name        Std.Dev.
##  X        (Intercept) 2.21
##  Residual             1.00
```

Try to fit glmer2 as a Bayesian model in Stan…

```
model{
  real lambda[N];
  // likelihood
  for (i in 1:N) {
    lambda[i] <- exp( a0[i] +
    a_roach*roach1[i] +
    a_tx*treatment[i] +
    a_senior*senior[i] +
    log_exposure[i] );
    y[i] ~ poisson(lambda[i]);
  }
  // priors
  for (i in 1:N) {
    eta0[i] ~ normal(0,tau0);
  }
  b0 ~ normal(0,10);
  a_roach  ~ normal(0,10);
  a_tx     ~ normal(0,10);
  a_senior ~ normal(0,10);
  tau0 ~ uniform(0,50);
}
```

- This model was producing many convergence and other errors
- Tried some things at https://mc-stan.org/misc/warnings.html and linked pages but it didn't seem to help

Diagnosing problems

- Try the suggestions at https://mc-stan.org/misc/warnings.html (and links from there)
- Look at shinystan
- Try initializing the model with the "simpler" MLE estimates

```
init.fcn  <- function() {
as.list(c(b0=1.08, a_roach=0.02, a_tx=-
0.82, a_senior=-1.04) + rnorm(4,0,0.1))
}

roach_results_1a <-
stan(fit=roach_model_1a,data=roach_data,
init=init.fcn,save_warmup=F)

launch_shinystan(roach_results_1a)
```

- Shinystan shows serious bimodality
  - Causing the convergence problems
- We could not have seen this with glmer
- Makes me not trust either
  - The Bayesian model
  - The glmer model
- Try a (much) simpler model in stan()

A simpler model in stan()

```
model{
  real lambda[N];

  // likelihood
  for (i in 1:N) {
    lambda[i] <- exp( a0[i] +
    a_tx*treatment[i] );
    y[i] ~ poisson(lambda[i]);
  }

  // priors
  for (i in 1:N) {
    eta0[i] ~ normal(0,tau0);
  }

  b0 ~ normal(0,10);

  a_tx     ~ normal(0,10);
  tau0 ~ uniform(0,50);

}
```

```
roach_model_1b <-
stan(file="roach1b.stan",
data=roach_data,chains=0)

roach_results_1b <-
stan(fit=roach_model_1b,
data=roach_data,
init=init.fcn,save_warmup=F)
launch_shinystan(roach_results_1b)

print(roach_results_1b,pars=c("b0","a_tx"
,"tau0"))
```

```
      mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
b0    1.40    0.03 0.30  0.78  1.20  1.41  1.60  1.97   133 1.01
a_tx -0.82    0.04 0.38 -1.59 -1.07 -0.83 -0.58 -0.07   102 1.02
tau0  2.78    0.01 0.19  2.44  2.64  2.77  2.90  3.18   174 1.03
```

There's still some evidence of excess zeros in the data.

- This can be handled with a
  - Zero-inflated Poisson model
  - Hurdle model
- They do similar things, differ only in how the "excess zeros" are modeled.
- Some additional exploration of stan() models with this data suggests that a model with both a random effect and zero inflation won't work.
- We will drop the random effect to see how zero inflation works

glmer() can't do this!)

Zero-inflated Models…

- Zero-inflated model add a point-mass at zero to a Poisson distribution

$$p(y|\theta,\lambda) = \begin{cases} \theta + (1-\theta)\cdot e^{-\lambda} & \text{if } y_n = 0, \text{ and} \\ (1-\theta)\cdot \frac{\lambda^y}{y!} e^{-\lambda} & \text{if } y_n > 0. \end{cases}$$

  - Flip a coin with $P[Heads] = \theta$:
    - if the coin comes up heads, $y = 0$;
    - if the coin comes up tails, $y \sim Poisson(\lambda)$.
- This can only add more zero observations, since either the data is Poisson, or it is just zero (which would be extra zeros).

Nice reference:
https://mc-stan.org/docs/2_20/stan-users-guide/zero-inflated-section.html

Hurdle Models…

- Hurdle models simply adjust $P[y = 0]$ to some other value than the Poisson value $e^{-\lambda}$:

$$p(y|\theta,\lambda) = \begin{cases} \theta & \text{if } y = 0, \text{ and} \\ (1-\theta)\ \dfrac{\frac{\lambda^y}{y!}e^{-\lambda}}{1-e^{-\lambda}} & \text{if } y > 0 \end{cases}$$

  - If $\theta = e^{-\lambda}$, we get the Poisson distribution back
  - If $\theta < e^{-\lambda}$, we get a distribution with less zeros than the Poisson
  - If $\theta > e^{-\lambda}$, we get a distribution with more zeros than the Poisson

Nice reference:
https://mc-stan.org/docs/2_20/stan-users-guide/zero-inflated-section.html

Our model is a Zero-Inflated Model:

Let

- $z_i = 1$ if the building is roach-free post-experiment, 0 otherwise,; and
- $w_i = 1$ if the building was roach-free pre-experiment, 0 otherwise.

Then our model will be

$$
\begin{aligned}
z_i &\sim Bernoulli(p_i) \\
& \quad logit(p_i) = c_0 + c_1 \cdot w_i + c_2 \cdot tx_i \\
\text{If } z_i = 1, & \quad y_i \equiv 0 \\
\text{If } z_i = 0, & \quad y_i \sim Poisson(\lambda_i) \\
& \quad \log(\lambda_i) = a_0 + a_1 tx_i
\end{aligned}
$$

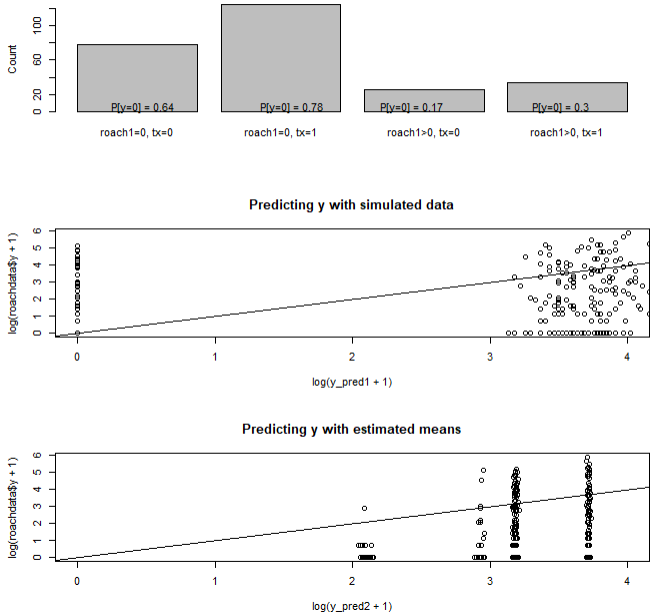
with prior distributions on c_0, c_1, c_2, $a0$ and a_1.

The model itself is a bit complicated in stan…

```
model{
  real p[N];
  real lambda[N];
  # likelihood
  for (i in 1:N) {
    lambda[i] <- exp( a0 + a_tx*treatment[i] );
    p[i] <- inv_logit(  c0 + c_roach*roach0[i] +
      c_tx*treatment[i]);
    target += if_else(z[i], log(p[i] + (1-p[i])*exp(-lambda[i])),
                            log(1-p[i]) + poisson_lpmf(y[i] | lambda[i]));
}
  # priors
  a0 ~ normal(0,10);
  a_tx     ~ normal(0,10);
  c0 ~      normal(0,10);
  c_roach ~ normal(0,10);
  c_tx ~    normal(0,10);
}
```

Amazingly it runs well in stan!

```
roach_model_zero_inf <-
stan(file="roach-zero-inflated.stan",

data=roach_data,chains=0)


roach_results_zero_inf <-
stan(fit=roach_model_zero_inf,
  data=roach_data,
  init=init.fcn,save_warmup=F)


launch_shinystan(roach_results_zero_inf)

print(roach_results_zero_inf,
pars=c("c0","c_roach","c_tx","a0","a_tx")
)

## stuff to make graphs of p's
## and of predicted vs actual y's
```



- Interesting model
- Doesn't fit as well as our Poisson model with a random intercept…

Summary

- Vignette 1: Meta Analysis
  - Example: Diet and Exercise vs Cholesterol
- Vignette 2: Multilevel Poisson regression and extensions
  - Example: the roaches data