Precept 4 - More GLMs: Models of Binary Outcomes

Soc 504: Advanced Social Statistics

Ian Lundberg[1]

Princeton University

March 2, 2018

[1]These slides owe an enormous debt to generations of TFs in Gov 2001 at Harvard. Many slides are directly adapted from those by Brandon Stewart and Stephen Pettigrew.

Outline

Replication Paper

Anything to discuss?

To discuss from a card

I would love it if the preceptors would be really honest about priorities in the homework. How should we prioritize the different elements each week? Especially weeks we're drowning other work? (1) Pset, (2) Replication? Reading?

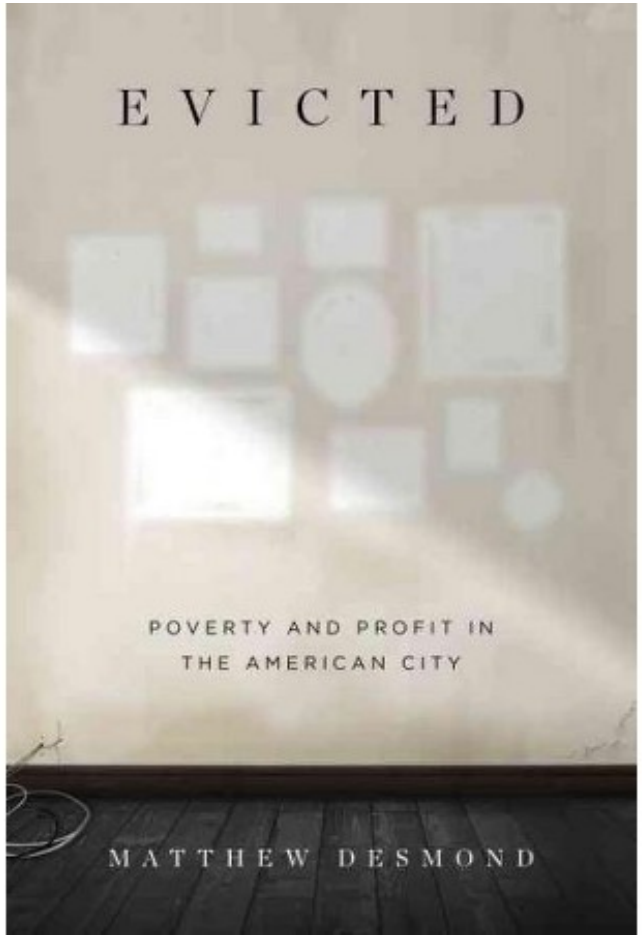
A framework for doing research

This should feel automatic

This is social science!

1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!


EVICTED
POVERTY AND PROFIT IN
THE AMERICAN CITY
MATTHEW DESMOND

We will use data from the Fragile Families and Child Wellbeing Study to study the probability of eviction for children born in large American cities.

Fragile Families
& Child Wellbeing Study
PRINCETON | COLUMBIA

Fragile Families data

`ffEviction.dta` contains these data.

`evicted`: was this child evicted in a given year?

`income`: family income / poverty line at age 1

`married`: were the parents married at the birth?

`race`: mother's race/ethnicity

`m1natwt`: sample weight

Fragile Families data

	idnum	evicted	income	married	race	m1natwt
1	0001	FALSE	1.5	FALSE	Hispanic	5.06258
2	0003	FALSE	2.7	FALSE	White	12.91446
3	0004	FALSE	1.0	FALSE	Hispanic	36.08243
4	0006	FALSE	0.2	FALSE	Black	79.70869
5	0007	FALSE	1.3	FALSE	Hispanic	31.66235
6	0008	FALSE	0.5	FALSE	Black	65.91409

A framework for doing research

This should feel automatic

This is social science!

1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

1. Define a quantity of interest τ

Note: I use the general notation (i.e. τ, θ) in the slide titles and the specific application notation in the slide text. τ represents a generic quantity of interest; π represents the particular quantity of interest in this example. Fitting a particular example into a general framework is a key goal of the course!

$\pi =$ probability of eviction in a calendar year for

- a white child
- born in 1998-2000
- to married parents
- living at the poverty line
- in an American city with population over 200,000

Research question: What is the probability of eviction in a calendar year for a white child born in 1998-2000 to married parents living at the poverty line in an American city with population over 200,000?

A simple strategy:

- There are 3,442 observations in the data.
- There are 7 observations meeting this description.
- One could report the proportion of these who are evicted.

Think, pair, share:

- **Q:** What are the advantages of this model-free approach?
  - **A:** It is unbiased, simple to understand, and requires minimal assumptions!
- **Q:** Why might one prefer a parametric model?
  - **A:** The model-free approach is very noisy. There are only 7 observations! We gain efficiency by assuming a parametric model in order to **share information** from other observations.

A framework for doing research

This should feel automatic

This is social science!

1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

Generalized Linear Models: Three elements

Linear predictor

$$\overbrace{\vec{X}_i\vec{\beta} = \eta_i}$$

Link function g

$$\overbrace{\eta_i = g(\mu_i)}$$

Stochastic component

$$\overbrace{Y_i \sim f_Y(\mu_i, \gamma)}$$

Specify a distribution for Y

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

Range is $[0, 1]$

Specify a linear predictor:

$$\eta_i = \vec{X}_i \vec{\beta}$$

Range is $(-\infty, \infty)$

Specify a link function $\eta_i = g(\pi_i)$

Complementary Log-log (cloglog):

$$\eta_i = g(\pi_i) = \underbrace{\log(\overbrace{-\log(1-\pi_i)}^{\text{Range is } (0,\infty)})}_{\text{Range is } (-\infty,\infty)}$$

Our complementary log-log link function is

$$\vec{X}_i\vec{\beta} = \eta_i = g(\pi_i) = \underbrace{\log(\overbrace{-\log(1-\pi_i)}^{\text{Range is } (0,\infty)})}_{\text{Range is } (-\infty,\infty)}$$

We can solve for π_i to get the inverse link function

$$\begin{aligned}
\overbrace{\vec{X}_i\vec{\beta} = \log(-\log(1-\pi_i))}^{\text{Link function } g} \\
\exp(\vec{X}_i\vec{\beta}) = -\log(1-\pi_i) \\
\exp(-\exp(\vec{X}_i\vec{\beta})) = 1-\pi_i \\
\underbrace{1-\exp(-\exp(\vec{X}_i\vec{\beta}))}_{\text{Inverse link function } g^{-1}} = \pi_i
\end{aligned}$$

There are many link functions for binary outcomes:

- Complementary log-log: $\eta_i = g(\pi_i) = \log(-\log(1-\pi_i))$
- Probit: $\eta_i = g(\pi_i) = \Phi^{-1}(\pi_i)$
- Logit: $\eta_i = g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$

Log-likelihood of the c-loglog

$$
\begin{aligned}
\ell(\vec{\beta} \mid \vec{Y}) &= \log(L(\vec{\beta} \mid \vec{Y})) \\
&= \log(p(\vec{Y} \mid \vec{\beta}, \mathbf{X})) \\
&= \log\left(\prod_{i=1}^{n} p(Y_i \mid \vec{\beta}, \vec{X}_i)\right) \qquad \leftarrow \text{ assumes independence conditional on } \mathbf{X} \\
&= \log\left(\prod_{i=1}^{n} [1 - \exp(-\exp(\vec{X}_i\vec{\beta}))]^{Y_i} [\exp(-\exp(\vec{X}_i\vec{\beta}))]^{(1-Y_i)}\right) \\
&= \sum_{i=1}^{n} \left(Y_i \log(1 - \exp(-\exp(\vec{X}_i\vec{\beta}))) + (1 - Y_i) \log[\exp(-\exp(\vec{X}_i\vec{\beta}))]\right) \\
&= \sum_{i=1}^{n} \left(Y_i \log(1 - \exp(-\exp(\vec{X}_i\vec{\beta}))) - (1 - Y_i) \exp(\vec{X}_i\vec{\beta})\right)
\end{aligned}
$$

Coding our log likelihood function

$$\sum_{i=1}^{n} \left(Y_i \log(1 - \exp(-\exp(\vec{X}_i \vec{\beta}))) - (1 - Y_i) \exp(\vec{X}_i \vec{\beta}) \right)$$

```
cloglog.loglik <- function(par, X, y) {

    beta <- par

    log.lik <- sum(y * log(1 - exp(-exp(X %*% beta))) -
                   (1 - y) * exp(X %*% beta))

  return(log.lik)
}
```

Finding the MLE

```
X <- model.matrix(~married + race + income,
                  data = d)

opt <- optim(par = rep(0, ncol(X)),

             fn = cloglog.loglik,

             X = X,
             y = d$ev,

             control = list(fnscale = -1),

             hessian = T,

             method = "BFGS")
```

Point estimate of the MLE:

```
opt$par
[1] -2.9980367 -1.0960323 -0.1799426  0.2055103  0.5671903
```

Standard errors of the MLE

Recall that the standard errors are defined as the diagonal of:

$$\sqrt{-\left[\frac{\partial^2 \ell}{\partial \beta^2}\right]^{-1}}$$

where $\frac{\partial^2 \ell}{\partial \beta^2}$ is the

2

[2] Credit to Stephen Pettigrew for including this figure in slides.

Standard errors of the MLE

Variance-covariance matrix:

```
-solve(opt$hessian)
```

	(Intercept)	married	raceBlack	raceHispanic	raceOther	income
(Intercept)	0.08	-0.01	-0.06	-0.06	-0.05	-0.01
married	-0.01	0.14	0.01	0.01	-0.00	-0.01
raceBlack	-0.06	0.01	0.08	0.06	0.05	0.01
raceHispanic	-0.06	0.01	0.06	0.08	0.05	0.01
raceOther	-0.05	-0.00	0.05	0.05	0.22	0.00
income	-0.01	-0.01	0.01	0.01	0.00	0.01

Standard errors: Square root of the diagonal

```
sqrt(diag(-solve(opt$hessian)))
```

```
 (Intercept)      married    raceBlack raceHispanic    raceOther       income
       0.276        0.370        0.281        0.280        0.466        0.088
```

Interpreting c-loglog coefficients

Here's a nicely formatted table with your regression results from our model:

Variable	Coefficient	SE
Intercept	-3.00	0.28
Married	-1.10	0.37
Black	-0.18	0.28
Hispanic	0.21	0.28
Other	0.57	0.47
Income / poverty line	-0.18	0.09

But what does this table tell us?

Interpreting c-loglog results

What does it mean for the coefficient for married to be -1.10?

All else constant, children of married parents have -1.10 points lower log rate of eviction.

Was that the kind of question that inspired you to become a social scientist? What are log rates? Nobody thinks in terms of log odds, or probit coefficients, or exponential rates.

If there's one thing you take away from this class, it should be this:

When you present results, **always** present your findings in terms of something that has substantive meaning to the reader.

For binary outcome models that often means turning your results into predicted probabilities, which is what we'll do now.

If there's a second thing you should take away, it's this:

Always account for all types of uncertainty when you present your results

We'll spend the rest of today looking at how to do that.

A framework for doing research

This should feel automatic

This is social science!

1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

3. Calculate the MLE for our quantity of interest $\tau = h(\theta)$

Our complementary log-log link function is

$$g(\pi_i) = \underbrace{\log(\overbrace{-\log(1-\pi_i)}^{\text{Range is } (0,\infty)})}_{\text{Range is } (-\infty,\infty)} \qquad g(\pi_i) = X_i\beta$$

To go from a covariate set X_i to a predicted probability, we use the inverse link function

$$\vec{X}_i\vec{\beta} = \overbrace{\log(-\log(1-\pi_i))}^{\text{Link function } g}$$

$$\underbrace{1-\exp(-\exp(\vec{X}_i\vec{\beta}))}_{\text{Inverse link function } g^{-1}(X_i\beta)} = \pi_i$$

The predicted probability of eviction for a child with covariates $\vec{x}$ is

$$\pi = h(\vec{x},\vec{\beta}) = g^{-1}(\vec{x}\vec{\beta}) = 1-\exp(-\exp(\vec{x}\vec{\beta}))$$

Calculate the MLE for our quantity of interest $\tau = h(\theta)$

The predicted probability of eviction for a child with covariates $\vec{x}$ is

$$\pi = h\left(\vec{x}, \vec{\beta}\right) = g^{-1}\left(\vec{x}\vec{\beta}\right) = 1 - \exp\left(-\exp\left(\vec{x}\vec{\beta}\right)\right)$$

$$\hat{\pi}_{\text{MLE}} = h\left(\vec{x}, \hat{\vec{\beta}}_{\text{MLE}}\right) = g^{-1}\left(\vec{x}\hat{\vec{\beta}}_{\text{MLE}}\right) = 1 - \exp\left(-\exp\left(\vec{x}\hat{\vec{\beta}}_{\text{MLE}}\right)\right)$$

In code:

```
get.pred.prob <- function(setX, beta) {
  ## Calculate the linear predictor
  eta <- setX %*% beta
  ## Transform by the inverse link function
  prob <- 1 - exp(-exp(eta))
  return(prob)
}
```

Calculate the MLE for our quantity of interest $\tau = h(\theta)$

Now we need to choose some values of the covariates that we want predictions about.

Let's make predictions for one white child born to married parents with family income at the poverty line Recall that our predictors (in order) are:

```
> colnames(X)
[1] "(Intercept)"        "married"           "cm1ethraceBlack"    "cm1ethraceHispanic"
[5] "cm1ethraceOther"    "income"
```

We can set the values of X as:

```
setX <- c(1, 1, 0, 0, 0, 1)
pi_hat <- get.pred.prob(setX = setX, beta = opt$par)
```

We estimate that the probability of eviction is $\hat{\pi} = 0.014$.

But how certain are we of that estimate?

A framework for doing research

This should feel automatic

This is social science!

1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

4. Simulate estimation uncertainty

We have estimation and fundamental uncertainty about Y.

In most models, we have to account for both types of uncertainty.

In this case, our quantity of interest $\pi = E(Y)$.

We are uncertain about π only because we are uncertain about β (estimation uncertainty).

Let's capture that uncertainty!

4. Simulate estimation uncertainty

$$\tilde{\beta} \sim \text{Normal}\left(\hat{\beta}, \hat{\text{V}}(\hat{\beta})\right)$$

$$\tilde{\pi} = h(x, \tilde{\beta})$$

Take one draw of $\tilde{\pi}$.

```
draw.sim.prob <- function(setX, beta_hat, vcov_beta_hat) {
  beta_tilde <- t(rmvnorm(n = 1,
                          mean = beta_hat,
                          sigma = vcov_beta_hat))
  prob_tilde <- get.pred.prob(setX = setX, beta = beta_tilde)
  return(prob_tilde)
}
```

4. Simulate estimation uncertainty

$$\tilde{\beta} \sim \text{Normal}\left(\hat{\beta}, \hat{\text{V}}(\hat{\beta})\right)$$

$$\tilde{\pi} = h(x, \tilde{\beta})$$

Take many draws of $\tilde{\pi}$.

```
beta_hat <- opt$par
vcov_beta_hat <- -solve(opt$hessian)
set.seed(08544)
draws <- replicate(20000, draw.sim.prob(setX = setX,
                                        beta_hat = beta_hat,
                                        vcov_beta_hat = vcov_beta_hat))
```

A framework for doing research

This should feel automatic

This is social science!

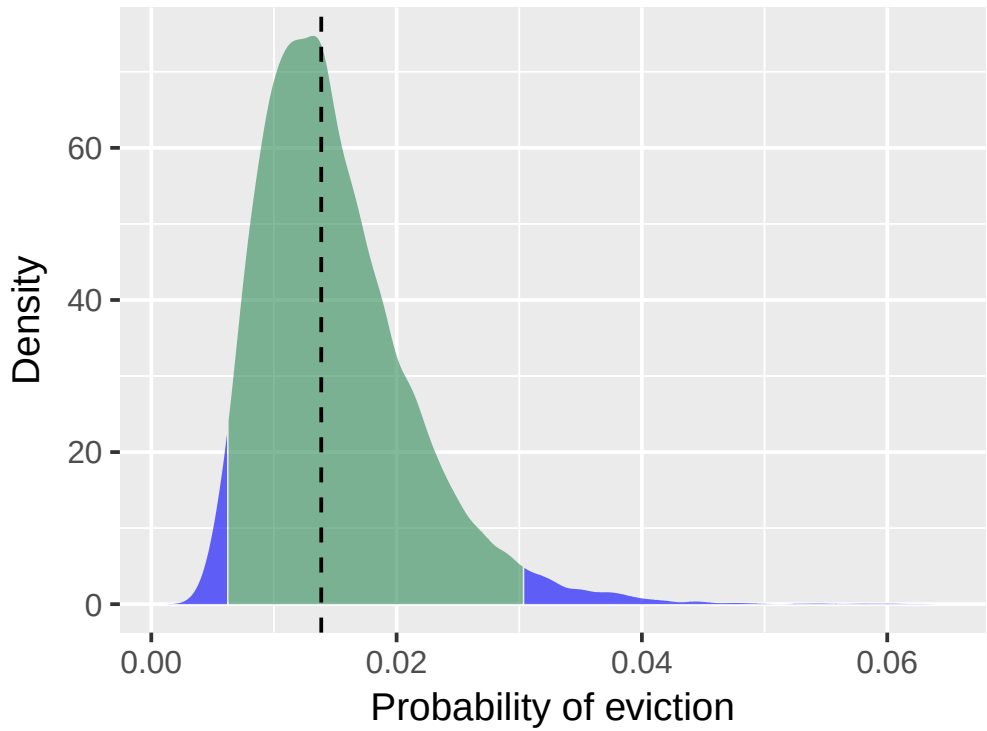
1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

5. Report results

We recommend reporting

- your point estimate $\hat{\tau}$ (the MLE by the invariance property)
- a 95% quantile-based confidence interval



A framework for doing research

This should feel automatic

This is social science!

1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

More practice.

Different quantity of interest

How does the probability of eviction vary by family income?
We will set $\vec{x}$ to represent a white child born to married parents and will vary the family income between twice the poverty line and half the poverty line.

$$\pi_{\text{Twice}} = \text{P}\left(Y \mid \vec{x}_{\text{Twice poverty line}}\right)$$
$$\pi_{\text{Half}} = \text{P}\left(Y \mid \vec{x}_{\text{Half poverty line}}\right)$$

$$\tau = \pi_{\text{Twice}} - \pi_{\text{Half}}$$

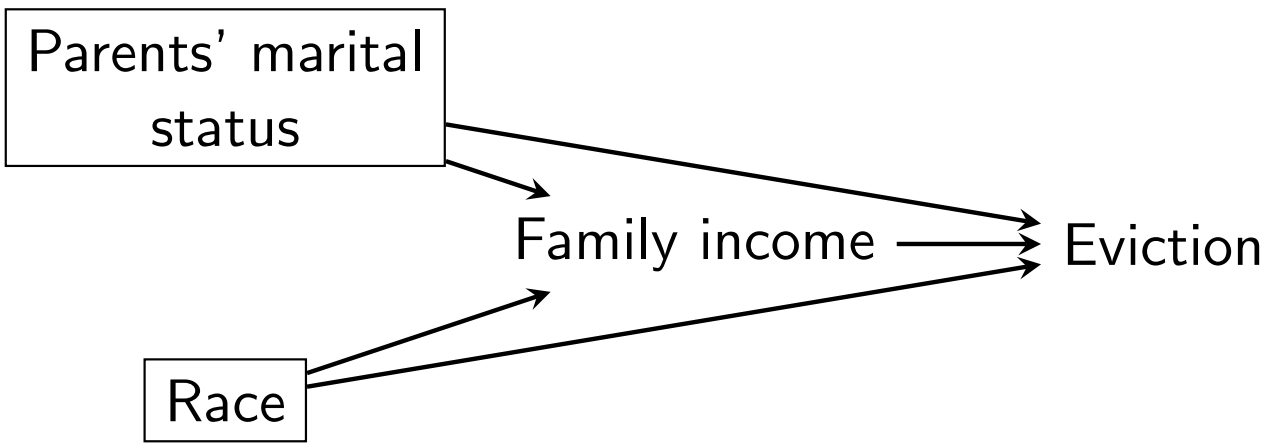
The model is the same. We just changed the quantity of interest.

Pause: Under what conditions is τ causal?

τ represents the difference in the predicted probability of eviction between children at twice the poverty line vs. half the poverty line, conditional on race and parents' marital status.

Q: What would we have to assume for this to be causal?

Conditional ignorability: No unblocked backdoor paths.



These are identification assumptions. We are also making estimation assumptions.
Often the quantity of interest is causal, but be careful to acknowledge the strong assumptions required!

3. Calculate the MLE for our quantity of interest $\tau = h(\theta)$

The true probability of eviction for a child with covariates $\vec{x}$ in a given year is

$$\tau = \pi_{\text{Twice}} - \pi_{\text{Half}}$$

1. Plug in the MLE estimates $\hat{\beta}$
2. Calculate $\hat{\pi}_{\text{Twice}}$ and $\hat{\pi}_{\text{Half}}$
3. Calculate $\hat{\tau}$

All works by the invariance property.

3. Calculate the MLE for our quantity of interest $\tau = h(\theta)$

In code:

```
> colnames(X)
[1] "(Intercept)"  "married"      "raceBlack"
     "raceHispanic" "raceOther"    "income"

setX <- rbind(deep_poverty = c(1, 1, 0, 0, 0, .5),
              twice_poverty = c(1, 1, 0, 0, 0, 2))

get.pred.diff <- function(setX, beta) {

  probs <- get.pred.prob(setX, beta)

  difference <- probs[2] - probs[1]

  return(difference)
}
```

A framework for doing research

This should feel automatic

This is social science!

1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

4. Simulate estimation uncertainty

$$\tilde{\beta} \sim \text{Normal}\left(\hat{\beta}, \hat{\text{V}}(\hat{\beta})\right)$$

$$\tilde{\pi} = h(x, \tilde{\beta})$$

$$\tilde{\tau} = \tilde{\pi}_{\text{Twice}} - \tilde{\pi}_{\text{Half}}$$

Take one draw of $\tilde{\tau}$.

```
draw.sim.diff <- function(setX, beta_hat, vcov_beta_hat) {
  beta_tilde <- t(rmvnorm(n = 1,
                          mean = beta_hat,
                          sigma = vcov_beta_hat))
  difference_tilde <- get.pred.diff(setX = setX, beta = beta_tilde)
  return(difference_tilde)
}
```

4. Simulate estimation uncertainty

Take many draws of $\tilde{\tau}$.

```
beta_hat <- opt$par
vcov_beta_hat <- -solve(opt$hessian)
set.seed(08544)
draws <- replicate(20000,
                   draw.sim.diff(setX = setX, beta_hat = beta_hat,
                                 vcov_beta_hat = vcov_beta_hat))
```

A framework for doing research

This should feel automatic

This is social science!

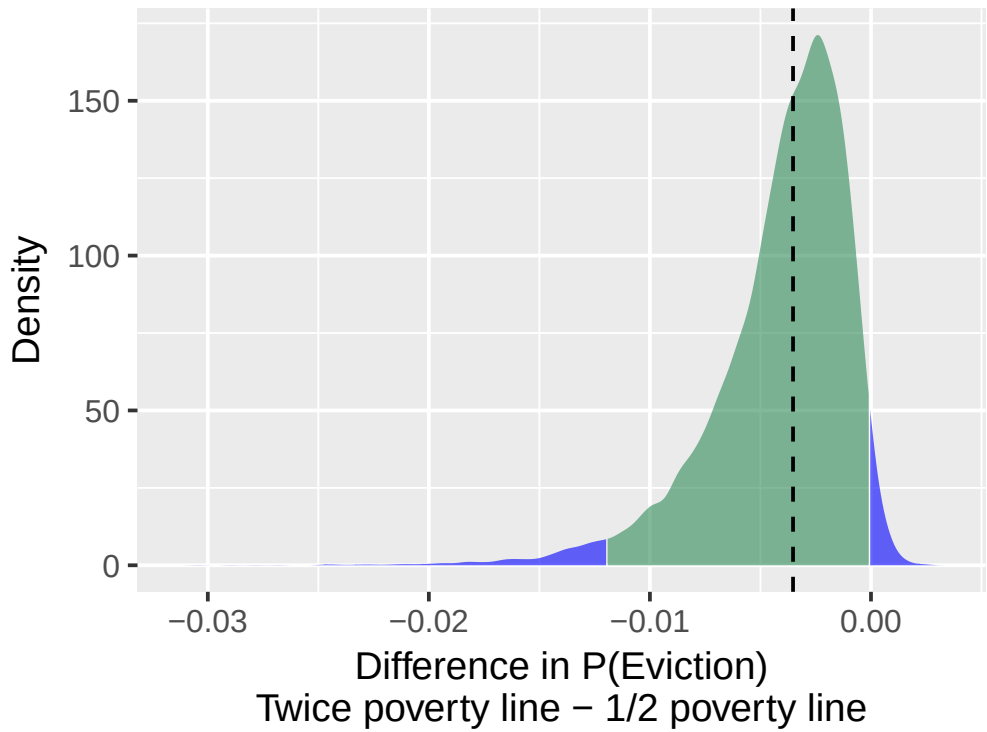
1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

5. Report results

We recommend reporting

- your point estimate $\hat{\tau}$ (the MLE by the invariance property)
- a 95% quantile-based confidence interval



A framework for doing research

This should feel automatic

This is social science!

1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

More practice.

Different quantity of interest

What is the probability of any eviction from birth to age 9, for a randomly sampled child born in a large American city?

$$\begin{aligned}\pi_i^{\text{Ever}} = \text{P}(\text{Ever evicted} \mid \vec{X}_i) &= 1 - \text{P}(\text{Never evicted} \mid \vec{X}_i) \\ &= 1 - \prod_{t=1}^{9}(1 - \text{P}(\text{Evicted at age } t \mid \vec{X}_i)) \\ &= 1 - (1 - \pi_i)^9\end{aligned}$$

The model is the same. We just changed the quantity of interest. [3]

[3] Note: We assume independence between eviction in each year, and a constant risk over time. This corresponds to an Exponential survival model.

Calculate the MLE for our quantity of interest $\tau = h(\theta)$

The true probability of eviction for a child with covariates $\vec{x}$ in a given year is

$$\pi_i = 1 - \exp(-\exp(\vec{X}_i\vec{\beta}))$$

The true probability of ever being evicted is:

$$\pi_i^{\text{Ever}} = h\left(\vec{x}, \vec{\beta}\right) = 1 - (1 - \pi_i)^9$$

We might want to report the weighted sample average of the π_i.

$$\bar{\pi}^{\text{Ever}} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i \pi_i^{\text{Ever}}$$

Plug in the MLE estimates $\hat{\beta}$ and solve! (invariance property).

Why weight? Assuming the model is correctly specified, this is representative of the probability of eviction by age 9 for a randomly sampled child born in a U.S. city with population over 200,000 in 1998-2000.

Calculate the MLE for our quantity of interest $\tau = h(\theta)$

In code:

```
get.pred.cum <- function(setX, beta, weights) {
  ## Calculate the linear predictor
  eta <- setX %*% beta
  ## Transform by the inverse link function
  prob_annual <- 1 - exp(-exp(eta))
  prob_cum <- 1 - (1 - prob_annual) ^ 9
  return(weighted.mean(prob_cum,
                       w = weights))
  return(prob_cum)
}
```

A framework for doing research

This should feel automatic

This is social science!

1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

4. Simulate estimation uncertainty

$$\tilde{\beta} \sim \text{Normal}\left(\hat{\beta}, \hat{\text{V}}(\hat{\beta})\right)$$

$$\tilde{\pi} = h(x, \tilde{\beta})$$

$$\tilde{\pi}_i^{\text{Ever}} = h\left(\vec{x}, \tilde{\beta}\right) = 1 - (1 - \tilde{\pi}_i)^9$$

$$\tilde{\bar{\pi}}^{\text{Ever}} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i \tilde{\pi}_i^{\text{Ever}}$$

Take one draw of $\tilde{\bar{\pi}}^{\text{Ever}}$.

```
draw.sim.cum <- function(setX, beta_hat, vcov_beta_hat, weights) {
  beta_tilde <- t(rmvnorm(n = 1,
                          mean = beta_hat,
                          sigma = vcov_beta_hat))
  cum_tilde <- get.pred.cum(setX = setX, beta = beta_tilde, weights = w
  return(cum_tilde)
}
```

4. Simulate estimation uncertainty

Take many draws of $\tilde{\tilde{\pi}}^{\text{Ever}}$.

```
beta_hat <- opt$par
vcov_beta_hat <- -solve(opt$hessian)
set.seed(08544)
draw.sim.cum(setX = X, beta_hat = opt$par,
             vcov_beta_hat = -solve(opt$hessian),
             weights = d$m1natwt)
```

A framework for doing research

This should feel automatic

This is social science!

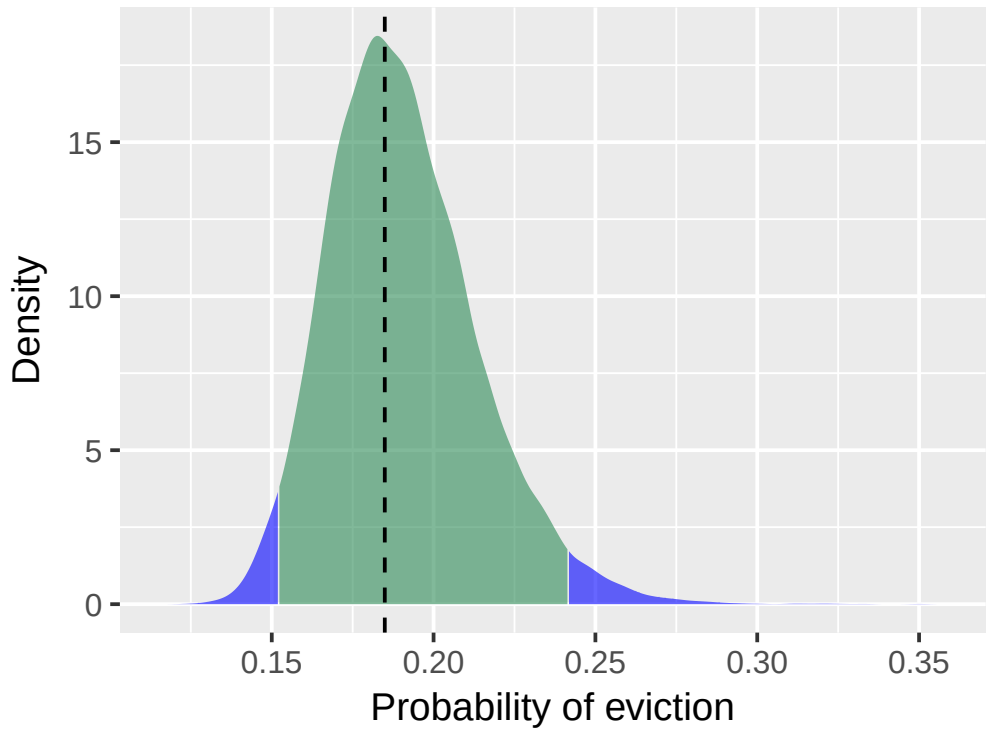
1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

Communication matters!

5. Report results

We recommend reporting

- your point estimate $\hat{\tau}$ (the MLE by the invariance property)
- a 95% quantile-based confidence interval



A framework for doing research

This should feel automatic

This is social science!

1. Define a quantity of interest τ
2. Specify a model and maximize the likelihood to estimate $\hat{\theta}$
3. Calculate the MLE for your quantity of interest: $\hat{\tau} = h(\hat{\theta})$
4. Simulate estimation uncertainty. Do the following thousands of times:
   1. Draw $\tilde{\theta}$ from its theoretical sampling distribution
   2. Calculate your quantity of interest $\tilde{\tau} = h\left(\tilde{\theta}\right)$
5. Report your point estimate from (3) and a 95% confidence interval from (4) in an informative graph.

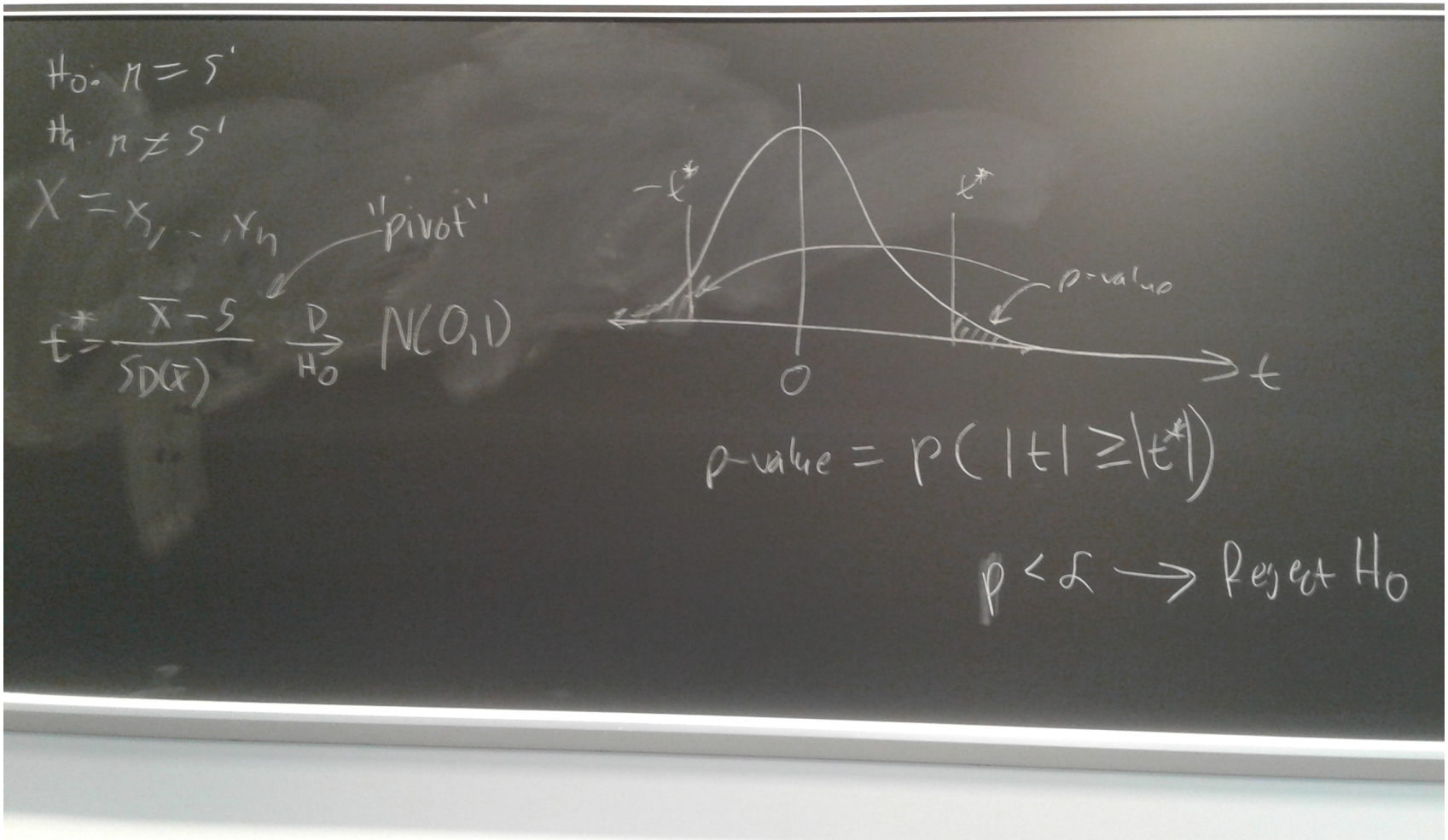
Communication matters!

Appendix: Hypothesis tests

We went through this quickly. Please ask questions on Piazza or in office hours and we will try to clarify! These last two slides have what we wrote on the board.

Hypothesis test for a mean

This should be review from the fall.



Likelihood ratio test

This is new this semester.