

STA2201H Methods of Applied Statistics II

Monica Alexander

Week 2: Generalized Linear Models

Announcements

- ▶ Piazza
- ▶ Assignment 1
- ▶ Office hours

Today:

- ▶ GLMs review (Counts and binary data)
- ▶ EDA in lab

GLMs

GLMs

Where we were at:

- ▶ General linear models (e.g. multivariate linear regression) not appropriate for some outcome variables
- ▶ In particular, when the outcome Y has a restricted range or the variance depends on the mean
- ▶ Generalized Linear Models extend the classical set-up to allow for a wider range of distributions
- ▶ GLMs have three pieces
 1. **random component:** $Y_i \sim$ some distribution with $E[Y_i | \mathbf{X}_i] = \mu_i$
 2. **systematic component:** $\mathbf{X}_i^T \beta$
 3. The **link function** that links the random and systematic components $g(u_i) = \mathbf{X}_i^T \beta$

What can Y be distributed as? In principle, anything. In practice (and original formulation), distributions come from the **exponential family**.

Exponential Family

Exponential Family

The random variable Y belongs to the exponential family of distributions if its support does not depend upon any unknown parameters and its density or probability mass function takes the form

$$p(y|\theta, \phi) = \exp \left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right)$$

- ▶ $\theta = h(\mu)$ depends on the expected value of y and is the **canonical parameter**
- ▶ ϕ is the scale parameter (if known: one-parameter family)
- ▶ b and c are arbitrary functions

Example: Poisson distribution

$$p(y|\theta, \phi) = \exp \left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right)$$

Poisson:

$$p(y|\mu) = \frac{\mu^y e^{-\mu}}{y!}$$

Write as

$$p(y|\mu) = \exp \{y \log \mu - \mu - \log y!\}$$

- ▶ $\theta = \log \mu$
- ▶ $b(\theta) = e^\theta$
- ▶ $c(y, \phi) = -\log y!$
- ▶ Note that the scale parameter $\phi = 1$ so the variance is entirely determined by the mean

Example: Normal distribution

$$p(y|\theta, \phi) = \exp \left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right)$$

Normal:

$$p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\}$$

Write as

$$p(y|\mu, \sigma^2) = \exp \left\{ \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2} \left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right] \right\}$$

- ▶ $\theta = \mu$
- ▶ $b(\theta) = \frac{1}{2}\theta^2$
- ▶ $\phi = \sigma^2$
- ▶ $c(y, \phi) = -\frac{1}{2} \left[\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2) \right]$

Properties of exponential families

Mean and variance for exponential families

$$p(y|\theta, \phi) = \exp \left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right)$$

It can be shown that

$$E(Y|\theta, \phi) = b'(\theta) = \mu$$

and

$$\text{Var}(Y|\theta, \phi) = \phi b''(\theta)$$

Note the variance of Y depends not only on the scale parameter but also on a function of the mean.

Examples:

$$E(Y|\theta, \phi) = b'(\theta)$$

and

$$\text{Var}(Y|\theta, \phi) = \phi b''(\theta)$$

- ▶ Poisson: $E(Y|\theta, \phi) = e^\theta = \mu$, $\text{Var}(Y|\theta, \phi) = 1 \times e^\theta = \mu$
- ▶ Normal: $E(Y|\theta, \phi) = \theta = \mu$, $\text{Var}(Y|\theta, \phi) = \sigma^2 \times 1 = \sigma^2$

The canonical link

The link function $\eta_i = g(\mu)$ could in theory be any function linking the linear predictor to the distribution of the outcome variable, which is also is **monotonic** and **smooth**.

Recall $\theta = h(\mu)$. If we choose $g = h$, then

$$\theta_i = h(\mu_i) = h(h^{-1}(\eta_i)) = \eta_i = \mathbf{x}_i^T \beta$$

In other words, it ensures that the systematic component of our model is modeling the parameter of interest.

Canonical links

- ▶ Normal: identity $\theta = h(\mu) = \mu$
- ▶ Poisson: $\theta = h(\mu) = \log \mu$
- ▶ Binomial: $\theta = h(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
- ▶ Exponential/Gamma: $\theta = h(\mu) = -\mu^{-1}$
- ▶ Inverse Gaussian: $\theta = h(\mu) = \mu^{-2}$

Likelihood-based estimation and inference

Summary:

- ▶ Maximum likelihood estimation, similar to linear regression but has to be estimated iteratively (using Newton Raphson / Method of Scoring)
- ▶ Inference based on the limiting distribution for MLE

$$\hat{\beta} \sim N(\beta, I(\hat{\beta})^{-1})$$

Standard errors are the square roots of the inverse of the information matrix.

- ▶ Use this for the classic Wald Tests e.g. $\sqrt{W} = \frac{\hat{\beta} - \beta_0}{\text{se}(\hat{\beta})}$ follows z distribution.

Likelihood ratio test

Testing nested models, ω_1 and ω_2 , $\omega_1 \in \omega_2$ and number of parameters $p_2 > p_1$

$$2[\log \ell(\hat{\beta}_1|\mathbf{y}) - \log \ell(\hat{\beta}_2|\mathbf{y})] \sim \chi_{p_1-p_2}$$

- ▶ Comparing fit of two models
- ▶ Model with more predictors will almost always fit better, but is the difference significant?

Poisson regression

Review

- ▶ mean ?
- ▶ variance ?
- ▶ link: ?

What's a problem with just looking at counts?

Offsets

$$\begin{aligned} Y_i &\sim \text{Poisson}(\lambda_i) \\ \text{or } Y_i &\sim \text{Poisson}(\mu_i O_i) \\ \log \mu_i &= \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned}$$

Offset controls for exposure to risk/making inferences to some baseline. e.g.

- ▶ population size
- ▶ age
- ▶ time since exposed

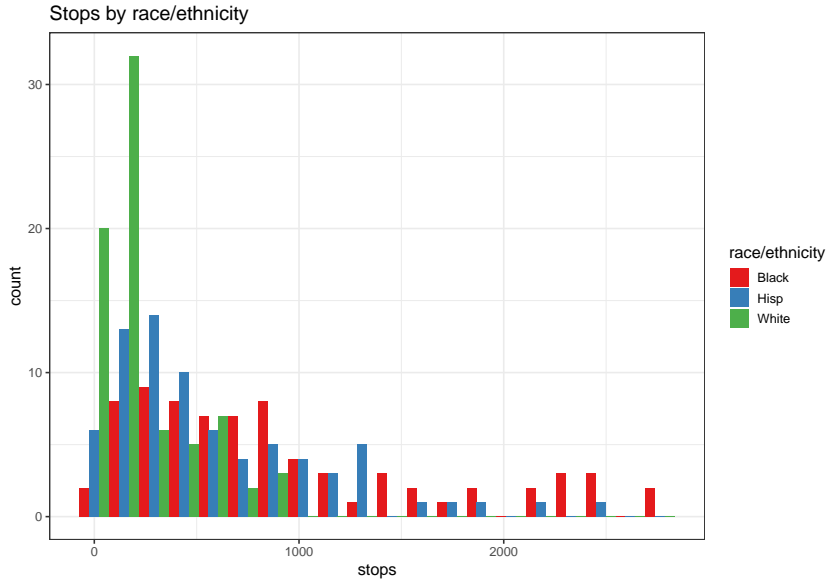
Example: Police stops

Police stop and frisks in NYC (Gelman Hill Chapter 6). Is there a difference in the number of stops by race/ethnicity?

The data look like:

precinct	stops	arrests	race_eth
1	202	980	Black
1	102	295	Hisp
1	81	381	White
2	132	753	Black
2	144	557	Hisp
2	71	431	White
3	752	2188	Black
3	441	627	Hisp
3	410	1238	White
4	385	471	Black

Distribution



Use arrests as exposure

```
mod1 <- glm(stops~race_eth,family=poisson,offset=log(arrests),data=d)
summary(mod1)
```

```
##
## Call:
## glm(formula = stops ~ race_eth, family = poisson, data = d, offset = log(arrests))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -47.327   -7.740   -0.182   10.241   39.140
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.588086   0.003784 -155.40  <2e-16 ***
## race_ethHisp   0.070208   0.006061  11.58  <2e-16 ***
## race_ethWhite -0.161581   0.008558  -18.88  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 46120  on 224  degrees of freedom
## Residual deviance: 45437  on 222  degrees of freedom
## AIC: 47150
##
## Number of Fisher Scoring iterations: 5
```

Add in factors for precinct

```
mod2 <- glm(stops~race_eth + factor(precinct), family=poisson, offset=log(arrests), data=d)
summary(mod2)[["coefficients"]][1:10,]
```

	Estimate	Std. Error	z value	Pr(> z)
## (Intercept)	-1.37886803	0.051019006	-27.026556	7.205634e-161
## race_ethHisp	0.01018798	0.006802045	1.497782	1.341899e-01
## race_ethWhite	-0.41900122	0.009434996	-44.409261	0.000000e+00
## factor(precinct)2	-0.14904964	0.074030344	-2.013359	4.407691e-02
## factor(precinct)3	0.55995498	0.056758425	9.865583	5.869222e-23
## factor(precinct)4	1.21063605	0.057548994	21.036615	3.032678e-98
## factor(precinct)5	0.28286532	0.056794015	4.980548	6.340447e-07
## factor(precinct)6	1.14420375	0.058047383	19.711547	1.716374e-86
## factor(precinct)7	0.21817307	0.064335032	3.391202	6.958688e-04
## factor(precinct)8	-0.39056473	0.056867814	-6.867940	6.513564e-12

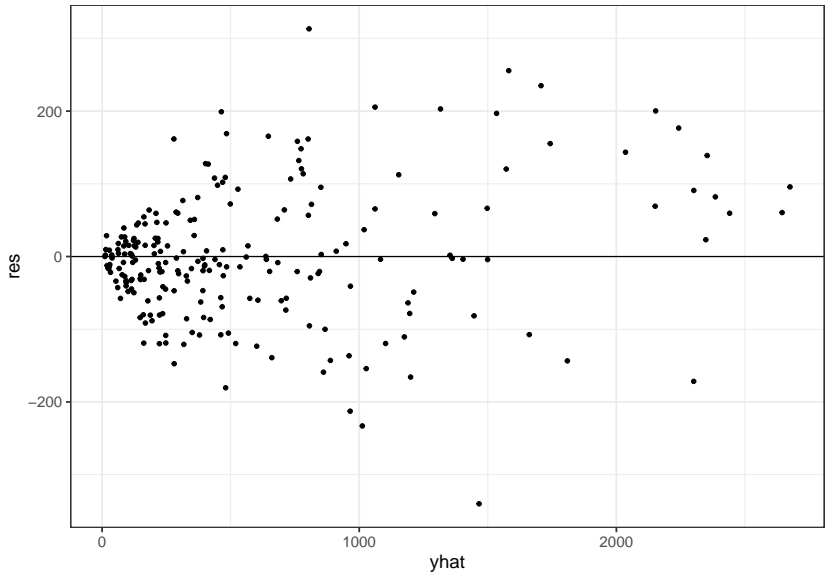
Coefficient interpretation

- ▶ e.g. after controlling for precinct, compared to blacks, whites have $1 - \exp(-0.42) = 34\%$ less chance of being stopped.
- ▶ be wary of exposure variable: stops are compared to the number of arrests in the previous year
- ▶ so that the coefficient 'whites' will be less than 1 if the people in that group are stopped disproportionately less than their rates of arrest, as compared to blacks.
- ▶ would be different if we had population as exposure variable

Is this a reasonable model?

Look at predicted values versus residuals $(y_i - \hat{y}_i)$. What do we expect?

Predicted values versus residuals



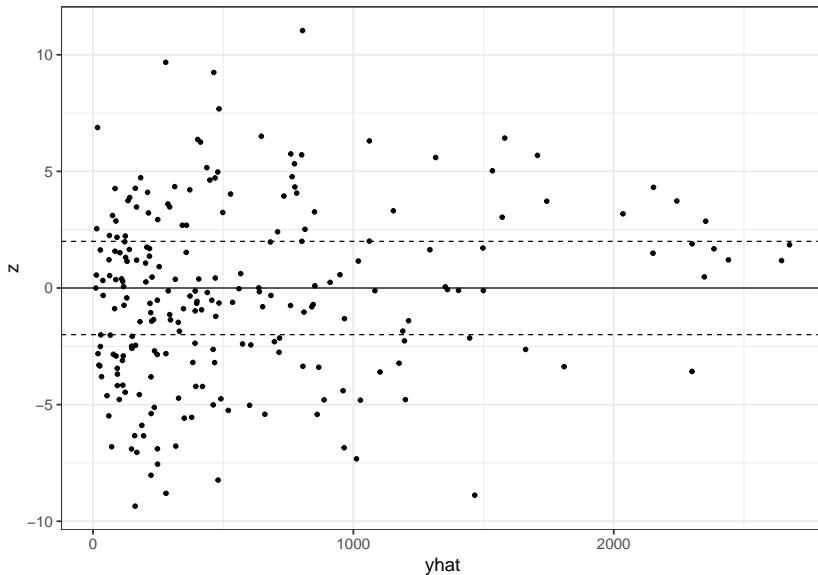
Is this a reasonable model?

Consider standardized residuals

$$z_i = \frac{y_i - \hat{y}_i}{sd(\hat{y}_i)}$$

If Poisson is a good model then these should have mean 0 and sd 1.

Predicted values versus standardized residuals



Overdispersion

- ▶ Extra variation in the data beyond what is allowed for in statistical model
- ▶ Poisson does not have independent variance parameter

Test for overdispersion: compare sum of squares of standardized residuals to χ^2_{n-k} distribution.

Estimated overdispersion factor is

$$\frac{1}{n-k} \sum_i z_i^2$$

Overdispersion

overdispersion factor is

```
sum(res_df$z^2)/(n-k)
```

```
## [1] 21.88505
```

Probability that we observe a factor at least as big as this is

```
1- pchisq(sum(res_df$z^2), n-k)
```

```
## [1] 0
```

But what's a problem here?

Fit overdispersed Poisson

- ▶ General form includes extra dispersion parameter θ
- ▶ Assume variance is proportion to the mean, rather than equal to the mean $E[Y] = \mu\theta$

```
mod3 <- glm(stops-race_eth + factor(precinct), family=quasipoisson,offset=log(arrests),data=d)
summary(mod3)[["coefficients"]][1:10,]
```

##		Estimate	Std. Error	t value	Pr(> t)
##	(Intercept)	-1.37886803	0.23867441	-5.7771925	4.326149e-08
##	race_ethHisp	0.01018798	0.03182097	0.3201657	7.492943e-01
##	race_ethWhite	-0.41900122	0.04413830	-9.4929170	5.489337e-17
##	factor(precinct)2	-0.14904964	0.34632483	-0.4303753	6.675488e-01
##	factor(precinct)3	0.55995498	0.26552425	2.1088656	3.664011e-02
##	factor(precinct)4	1.21063605	0.26922265	4.4967837	1.384310e-05
##	factor(precinct)5	0.28286532	0.26569075	1.0646412	2.887722e-01
##	factor(precinct)6	1.14420375	0.27155419	4.2135374	4.352372e-05
##	factor(precinct)7	0.21817307	0.30096874	0.7249028	4.696562e-01
##	factor(precinct)8	-0.39056473	0.26603599	-1.4680898	1.442019e-01

Notice

```
summary(mod3)[["dispersion"]]
```

```
## [1] 21.88506
```

... and the SEs are inflated $\sim \sqrt{21.9}$.

Overdispersion

Downside to quasi-Poisson it's not true MLE so you don't get likelihood etc to compare models.

Alternative:

- ▶ Could also add a multiplicative random effect θ to represent unobserved heterogeneity.
- ▶ Conditional distribution is Poisson $E[Y|\theta] \sim \text{Pois}(\mu\theta)$
- ▶ Leads to unconditional distribution being Negative Binomial distribution
- ▶ Can choose parameters so $E(Y) = \mu$ and $\text{Var}(Y) = \mu(1 + \sigma^2\mu)$

Overdispersion

Fit Negative Binomial

```
library(MASS)
mod4 <- glm.nb(stops~race_eth + factor(precinct), data = d)
summary(mod3)[["coefficients"]][1:10,]
```

	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	-1.37886803	0.23867441	-5.7771925	4.326149e-08
## race_ethHispanic	0.01018798	0.03182097	0.3201657	7.492943e-01
## race_ethWhite	-0.41900122	0.04413830	-9.4929170	5.489337e-17
## factor(precinct)2	-0.14904964	0.34632483	-0.4303753	6.675488e-01
## factor(precinct)3	0.55995498	0.26552425	2.1088656	3.664011e-02
## factor(precinct)4	1.21063605	0.26922265	4.4967837	1.384310e-05
## factor(precinct)5	0.28286532	0.26569075	1.0646412	2.887722e-01
## factor(precinct)6	1.14420375	0.27155419	4.2135374	4.352372e-05
## factor(precinct)7	0.21817307	0.30096874	0.7249028	4.696562e-01
## factor(precinct)8	-0.39056473	0.26603599	-1.4680898	1.442019e-01

Binary data

Binary Responses

We have n random variables Z_1, \dots, Z_n that are binary

$$Z_i = \begin{cases} 1 & \text{if outcome is a success} \\ 0 & \text{if outcome is a failure} \end{cases}$$

with

$$Pr(Z_1 = 1) = \pi_i$$

so

$$Pr(Z_1 = 0) = 1 - \pi_i$$

Logistic regression

We are interested in describing the probability of success π_i with a linear model

$$g(\pi_i) = \mathbf{x}^T \beta$$

The **canonical link** is the logistic function, so

$$\text{logit } \pi_i = \log \frac{\pi_i}{1 - \pi_i} = \mathbf{x}^T \beta$$

Latent variable formulation

$$\begin{aligned}y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\z_i &= X_i\beta + \epsilon_i \\ \epsilon_i &\sim f(.)\end{aligned}$$

Latent variable formulation

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases}$$

$$z_i = X_i\beta + \epsilon_i$$

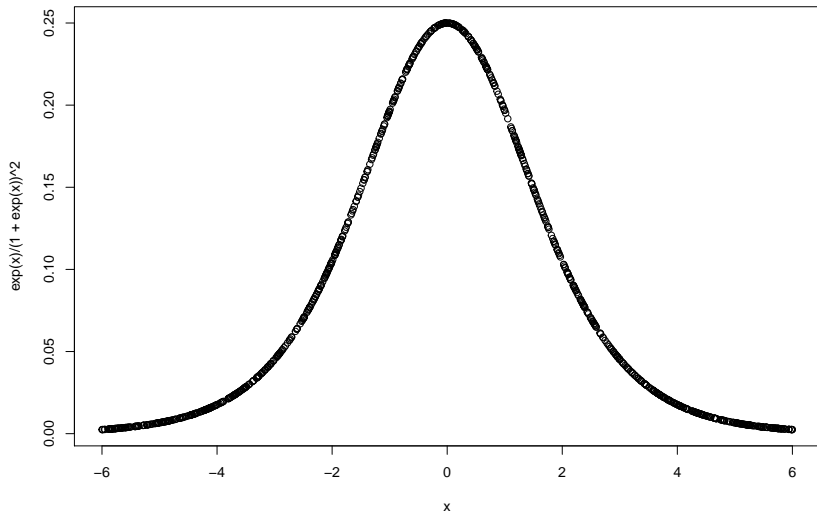
$$\epsilon_i \sim f(.)$$

For logistic regression, the errors ϵ have a *logistic* probability distribution

$$p(x) = \frac{e^x}{(1 + e^x)^2}$$

Latent variable formulation

The logistic pdf looks like



Latent variable formulation

Write $\eta_i = X_i\beta$.

Note that

$$\begin{aligned}\pi_i &= Pr(z_i > 0) \\ &= Pr(\epsilon_i > -\eta_i) \\ &= 1 - F(-\eta_i) \\ &= F(\eta_i)\end{aligned}$$

For the logistic, $F(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$ so $\eta_i = F^{-1}(\pi_i) = \frac{\pi_i}{1-\pi_i}$ as before.

Example: Abortion outcomes in Uganda

- ▶ Data from 2018 PMA survey (via IPUMS)
- ▶ Outcome of interest: 'ever had abortion (yes/no)'
- ▶ Other variables: age, region, urban/rural, education, marital status

What the data look like:

resp	urban	region	marstat	educattgen	abortion
1	rural	north	divorced or separated	primary/middle school	0
2	rural	eastern	currently married	primary/middle school	1
3	urban	kampala	divorced or separated	secondary/post-primary	0
4	rural	western	never married	secondary/post-primary	0
5	rural	karamoja	currently living with partner	never attended	0
6	rural	central 1	currently living with partner	secondary/post-primary	1

Abortion outcomes

Consider the model

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

$$\text{logit}\pi_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{marital}_i + \beta_3 \text{school}_i$$

where school_i is any level of schooling.

What could we use this model for? (i.e. what questions could we ask?)

Estimation in R

```
mod <- glm(abortion ~ age + marstat + school, data = d, family = "binomial")
summary(mod)
```

```
##
## Call:
## glm(formula = abortion ~ age + marstat + school, family = "binomial",
##      data = d)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.7456  -0.4965  -0.4364  -0.3433   2.5789
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.50979    0.26492  -13.248  < 2e-16 ***
## age              0.02562    0.00644   3.978 6.95e-05 ***
## marstatcurrently married -0.06072    0.12987  -0.468  0.64012
## marstatdivorced or separated 0.47349    0.15115   3.133  0.00173 **
## marstatnever married    -0.49844    0.22409  -2.224  0.02613 *
## marstatwidow or widower  -0.30133    0.36989  -0.815  0.41528
## school           0.84777    0.20125   4.212 2.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2564.4  on 3931  degrees of freedom
## Residual deviance: 2501.4  on 3925  degrees of freedom
## AIC: 2515.4
##
## Number of Fisher Scoring iterations: 5
```

Interpretation

```
coef(mod)
```

```
##                (Intercept)                age
##                -3.50979534                0.02561998
##   marstatcurrently married marstatdivorced or separated
##                -0.06072052                0.47348893
##   marstatnever married      marstatwidow or widower
##                -0.49844113                -0.30132839
##                school
##                0.84776732
```

```
exp(coef(mod))
```

```
##                (Intercept)                age
##                0.02990303                1.02595099
##   marstatcurrently married marstatdivorced or separated
##                0.94108622                1.60558621
##   marstatnever married      marstatwidow or widower
##                0.60747690                0.73983478
##                school
##                2.33442900
```

Questions

- ▶ What is the probability of ever had an abortion for a women aged 25, currently living with partner, who has never attended school?

```
estimated_log_odds <- coef(mod)[1] + coef(mod)[2]*25  
exp(estimated_log_odds)/(1+exp(estimated_log_odds))
```

```
## (Intercept)  
## 0.05369242
```

Questions

- ▶ Assume we don't observe women in a particular region. Could we use this model to predict the likelihood of abortion in this region?
- ▶ Can we use this model to estimate the impact of education on abortion?

What are some potential issues with this analysis?

$$\text{logit}\pi_i = \beta_0 + \beta_1\text{age}_i + \beta_2\text{marital}_i + \beta_3\text{school}_i$$

Model issues

- ▶ Omitted variable bias
- ▶ Model mis-specification
- ▶ Model underfit or overfit
- ▶ Multicollinearity

Tools (for now)

- ▶ EDA!
- ▶ Likelihood ratio tests
- ▶ Wald tests
- ▶ Assessing predictions/residuals graphically (harder with binary variables)

A good way of assessing model fit

Look at predicted v actual proportions by groups



Issues with causal questions

Consider the situations where we are interested in the impact of education on abortion outcomes.

```
summary(mod)
```

```
##
## Call:
## glm(formula = abortion ~ age + marstat + school, family = "binomial",
##      data = d)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.7456  -0.4965  -0.4364  -0.3433   2.5789
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -3.50979    0.26492  -13.248  < 2e-16 ***
## age             0.02562    0.00644   3.978  6.95e-05 ***
## marstatcurrently married -0.06072    0.12987  -0.468  0.64012
## marstatdivorced or separated 0.47349    0.15115   3.133  0.00173 **
## marstatnever married  -0.49844    0.22409  -2.224  0.02613 *
## marstatwidow or widower  -0.30133    0.36989  -0.815  0.41528
## school          0.84777    0.20125   4.212  2.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2564.4  on 3931  degrees of freedom
## Residual deviance: 2501.4  on 3925  degrees of freedom
## AIC: 2515.4
##
## Number of Fisher Scoring iterations: 5
```

Issues with causal questions

Consider the situations where we are interested in the impact of education on abortion outcomes. This is a causal question. What are some issues that may arise?

Issues with causal questions

- ▶ Confounders
 - ▶ urbanity
- ▶ Colliders (e.g. non-reponse bias)
 - ▶ Schooling and abortion both influence survey response
 - ▶ Conditioning on survey response creates a noncausal association between schooling and abortion

Data issues

- ▶ Non-representative samples
- ▶ Non-response (complete survey or specific questions)
- ▶ Measurement error
 - ▶ in this case, self-reports are a bad survey instrument

Multinomial

- ▶ Next week or could add additional notes

Lab

- ▶ Using data from Open Data Portal in Toronto
 - ▶ `opendatatoronto` package
- ▶ EDA
- ▶ Questions at end need to be handed in via GitHub
- ▶ Ask about `ggplot`?