MA678 Homework 5

Chang Lu

10/25/2022

15.1 Poisson and negative binomial regression

The folder `RiskyBehavior` contains data from a randomized trial targeting couples at high risk of HIV infection. The intervention provided counseling sessions regarding practices that could reduce their likelihood of contracting HIV. Couples were randomized either to a control group, a group in which just the woman participated, or a group in which both members of the couple participated. One of the outcomes examined after three months was "number of unprotected sex acts."

a)

Model this outcome as a function of treatment assignment using a Poisson regression. Does the model fit well? Is there evidence of overdispersion?

```
library(MASS)
```

```
data <- read.csv("risky.csv")
data$fupacts <- round(data$fupacts)

poisson_model <- glm(fupacts ~ couples + women_alone, family = poisson, data = data)

summary(poisson_model)
```

```
##
## Call:
## glm(formula = fupacts ~ couples + women_alone, family = poisson,
##     data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.08960    0.01901  162.55   <2e-16 ***
## couples     -0.32243    0.02737  -11.78   <2e-16 ***
## women_alone -0.57212    0.03023  -18.93   <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: 13299  on 433  degrees of freedom
## Residual deviance: 12925  on 431  degrees of freedom
## AIC: 14256
```

```
##
## Number of Fisher Scoring iterations: 6
```

```
# Check for overdispersion
dispersion <- sum(residuals(poisson_model, type = "deviance")^2) / poisson_model$df.residual
dispersion
```

```
## [1] 29.9895
```

No, the model doesn't fit well. For the calculated dispersion coefficient is 29.99, which is much greater than 1. This indicates strong evidence of overdispersion.

b)

Next extend the model to include pre-treatment measures of the outcome and the additional pre-treatment variables included in the dataset. Does the model fit well? Is there evidence of overdispersion?

```
extended_poisson_model <- glm(fupacts ~ couples + women_alone + bupacts + bs_hiv, family = poisson, data

summary(extended_poisson_model)
```

```
##
## Call:
## glm(formula = fupacts ~ couples + women_alone + bupacts + bs_hiv,
##     family = poisson, data = data)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.8415576  0.0201490  141.03   <2e-16 ***
## couples        -0.4127367  0.0282749  -14.60   <2e-16 ***
## women_alone    -0.6556596  0.0308079  -21.28   <2e-16 ***
## bupacts         0.0107541  0.0001741   61.78   <2e-16 ***
## bs_hivpositive -0.4331504  0.0353793  -12.24   <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: 13299  on 433  degrees of freedom
## Residual deviance: 10221  on 429  degrees of freedom
## AIC: 11556
##
## Number of Fisher Scoring iterations: 6
```

```
dispersion1 <-  sum(residuals(extended_poisson_model, type = "deviance")^2) / extended_poisson_model$df

dispersion1
```

```
## [1] 23.82599
```

No. The calculated dispersion coeffcient is still greater than 1, indicating overdispersion.

c)

Fit a negative binomial (overdispersed Poisson) model. What do you conclude regarding effectiveness of the intervention?

```
neg_binom_model <- glm.nb(fupacts ~ couples + women_alone + bupacts + bs_hiv, data = data)


summary(neg_binom_model)
```

```
##
## Call:
## glm.nb(formula = fupacts ~ couples + women_alone + bupacts +
##     bs_hiv, data = data, init.theta = 0.4168882054, link = log)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.450274   0.154745  15.834  < 2e-16 ***
## couples        -0.349628   0.188987  -1.850 0.064311 .
## women_alone    -0.715797   0.192398  -3.720 0.000199 ***
## bupacts         0.022374   0.002362   9.471  < 2e-16 ***
## bs_hivpositive -0.551637   0.185920  -2.967 0.003007 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.4169) family taken to be 1)
##
##     Null deviance: 581.84  on 433  degrees of freedom
## Residual deviance: 487.67  on 429  degrees of freedom
## AIC: 2967.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.4169
##           Std. Err.:  0.0313
##
##  2 x log-likelihood:  -2955.7500
```

The intervention where only the woman participated in the counseling sessions significantly reduced the number of unprotected sex acts ($p = 0.000199$).

The intervention where both partners participated (couples) showed a reduction in unprotected sex acts, but this was not statistically significant ($p = 0.064$).

Participants who were HIV-positive or had more unprotected sex acts before the intervention also exhibited notable differences in behavior post-intervention.

d)

These data include responses from both men and women from the participating couples. Does this give you any concern with regard to our modeling assumptions? In regression models (including Poisson and negative binomial models), one of the assumptions is that the observations are independent of each other. However,

in this dataset, men and women from the same couple are likely to influence each other's behaviors (e.g., the number of unprotected sex acts). Therefore, responses from the same couple may be correlated, violating this independence assumption.

15.3 Binomial regression

Redo the basketball shooting example on page 270, making some changes:

(a)

Instead of having each player shoot 20 times, let the number of shots per player vary, drawn from the uniform distribution between 10 and 30.

```
set.seed(123)

N <- 100

height <- rnorm(N, 72, 3)

n <- round(runif(N, min = 10, max = 30))

p <- 0.4 + 0.1 * (height - 72) / 3

y <- rbinom(N, n, p)

data2 <- data.frame(n = n, y = y, height = height)

head(data2)
```

```
##    n  y   height
## 1 15  7 70.31857
## 2 29  5 71.30947
## 3 22 10 76.67612
## 4 20  9 72.21153
## 5 18  8 72.38786
## 6 28 16 77.14519
```

(b)

Instead of having the true probability of success be linear, have the true probability be a logistic function, set so that Pr(success) = 0.3 for a player who is 5'9" and 0.4 for a 6' tall player.

```
logistic <- function(x) {
  exp(x) / (1 + exp(x))
}

# Set the true probability of success using a logistic function
# Height of 69 inches corresponds to a probability of 0.3 and 72 inches to 0.4
logit_p <- -2 + 0.1 * (height - 69)
p <- logistic(logit_p)
```

```r
# Simulate the number of successes
y <- rbinom(N, n, p)

data3 <- data.frame(n = n, y = y, height = height)

fit <- glm(cbind(y, n - y) ~ height, family = binomial(link = "logit"), data = data3)

summary(fit)
```

```
##
## Call:
## glm(formula = cbind(y, n - y) ~ height, family = binomial(link = "logit"),
##     data = data3)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.31510    1.63391  -5.701 1.19e-08 ***
## height       0.10541    0.02241   4.703 2.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 122.61  on 99  degrees of freedom
## Residual deviance: 100.41  on 98  degrees of freedom
## AIC: 366.35
##
## Number of Fisher Scoring iterations: 4
```

15.7 Tobit model for mixed discrete/continuous data

Experimental data from the National Supported Work example are in the folder `Lalonde`. Use the treatment indicator and pre-treatment variables to predict post-treatment (1978) earnings using a Tobit model. Interpret the model coefficients.

```r
library("AER")
```

```
## Loading required package: car

## Loading required package: carData

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
library("haven")
NSW_dw_obs <- read_dta("NSW_dw_obs.dta")
head(NSW_dw_obs)
```

```
## # A tibble: 6 x 12
##     age  educ black married nodegree   re74   re75   re78  hisp sample treat
##   <dbl> <dbl> <dbl>   <dbl>    <dbl>  <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl>
## 1    42    16     0       1        0     0      0    100.     0      2     0
## 2    20    13     0       0        0  2367.  3317.  4794.     0      2     0
## 3    37    12     0       1        0 25862. 22782. 25565.     0      2     0
## 4    48    12     0       1        0 21591. 20839. 20551.     0      2     0
## 5    51    12     0       1        0 21395. 21575. 22784.     0      2     0
## 6    18    11     0       0        1  1311.  1456.  2157.     0      2     0
## # i 1 more variable: educ_cat4 <dbl>
```

```
tobit_model <- tobit(re78 ~ treat + age + educ + re74 + re75, data = NSW_dw_obs, left = 0)

summary(tobit_model)
```

```
##
## Call:
## tobit(formula = re78 ~ treat + age + educ + re74 + re75, left = 0,
##     data = NSW_dw_obs)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##          18667           2503          16164              0
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  6.004e+03  3.602e+02   16.669  < 2e-16 ***
## treat        4.859e+02  6.543e+02    0.743    0.458
## age         -1.595e+02  6.491e+00  -24.567  < 2e-16 ***
## educ         1.059e+02  2.275e+01    4.655 3.25e-06 ***
## re74         3.365e-01  1.256e-02   26.789  < 2e-16 ***
## re75         5.680e-01  1.259e-02   45.098  < 2e-16 ***
## Log(scale)   9.046e+00  5.724e-03 1580.314  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 8487
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 5
## Log-likelihood: -1.717e+05 on 7 Df
## Wald-statistic: 1.766e+04 on 5 Df, p-value: < 2.22e-16
```

Coefficients

Intercept (6004):

The estimated latent earnings for a participant in the control group (no treatment), assuming all other variables are 0. This is not always a meaningful figure, since the value of the predictor variables would rarely all be zero in practice, but it serves as the baseline from which the other variables' effects are measured.

treat (485.9):

The treatment coefficient is 485.9, but it is not statistically significant (p = 0.458). This suggests that being in the treatment group did not significantly affect post-treatment earnings compared to the control group after controlling for other variables like age, education, and past earnings. age (-159.5):

For each additional year of age, post-treatment earnings decrease by 159.5 units (or dollars, assuming earnings are in dollars). This is statistically significant (p < 2e-16), meaning older participants generally earned less in 1978, controlling for other variables. educ (105.9):

Each additional year of education is associated with an increase in earnings of about 105.9 units. This effect is statistically significant (p = 3.25e-06), indicating that more years of education tend to increase post-treatment earnings. re74 (0.3365):

Earnings in 1974 have a significant positive effect on 1978 earnings. For every one-unit increase in 1974 earnings, 1978 earnings increase by 0.3365 units. This suggests a strong correlation between pre-treatment and post-treatment earnings, and it's highly significant (p < 2e-16). re75 (0.5680):

Similar to re74, earnings in 1975 have a highly significant positive effect on earnings in 1978. For every one-unit increase in 1975 earnings, 1978 earnings increase by 0.5680 units. This is also very significant (p < 2e-16). Log(scale) (9.046):

This is related to the standard deviation of the error term in the Tobit model. A larger value suggests more variation around the predicted values, and it is statistically significant, but it's more of a technical parameter than one of substantive interest.

Model Fit:

Scale (8487):

This is the estimated standard deviation (scale parameter) of the latent variable. It indicates that the model assumes considerable variability in earnings.

Log-likelihood (-171,700):

This is a measure of model fit, with larger (less negative) values indicating a better fit. It's useful for comparing different models, but on its own, it doesn't provide much insight.

Wald-statistic (17,660, p < 2.22e-16):

The Wald test assesses the overall significance of the model. The very high Wald-statistic and extremely small p-value suggest that the model as a whole is highly significant, meaning that at least one of the predictors significantly affects earnings in 1978.

15.8 Robust linear regression using the t model

The folder `Congress` has the votes for the Democratic and Republican candidates in each U.S. congressional district in 1988, along with the parties' vote proportions in 1986 and an indicator for whether the incumbent was running for reelection in 1988. For your analysis, just use the elections that were contested by both parties in both years.

```
library(rstanarm)
```

```
## Loading required package: Rcpp

## This is rstanarm version 2.32.1

## - See https://mc-stan.org/rstanarm/articles/priors for changes to default priors!

## - Default priors may change, so it's safest to specify priors, even if equivalent to the defaults.

## - For execution on a local, multicore CPU with excess RAM we recommend calling

##   options(mc.cores = parallel::detectCores())

##
## Attaching package: 'rstanarm'

## The following object is masked from 'package:car':
##
##     logit
```

```
library(brms)
```

```
## Loading 'brms' package (version 2.22.0). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').

##
## Attaching package: 'brms'

## The following objects are masked from 'package:rstanarm':
##
##     dirichlet, exponential, get_y, lasso, ngrps

## The following object is masked from 'package:survival':
##
##     kidney

## The following object is masked from 'package:stats':
##
##     ar
```

```
congress_data <- read.csv("congress.csv")
```

(a)

Fit a linear regression using `stan_glm` with the usual normal-distribution model for the errors predicting 1988 Democratic vote share from the other variables and assess model fit.

```r
# Fit a linear regression model using stan_glm
model_stan <- stan_glm(v88 ~ v86 + inc88, data = congress_data, family = gaussian, refresh = 0)

# Assess model fit
summary(model_stan)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      v88 ~ v86 + inc88
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 435
##  predictors:   3
##
## Estimates:
##               mean   sd   10%   50%   90%
## (Intercept) 0.3    0.0  0.3   0.3   0.3
## v86         0.4    0.0  0.4   0.4   0.5
## inc88       0.1    0.0  0.1   0.1   0.2
## sigma       0.1    0.0  0.1   0.1   0.1
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 0.6    0.0  0.6   0.6   0.6
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##               mcse Rhat n_eff
## (Intercept)   0.0  1.0  2230
## v86           0.0  1.0  2123
## inc88         0.0  1.0  2013
## sigma         0.0  1.0  2845
## mean_PPD      0.0  1.0  3747
## log-posterior 0.0  1.0  1790
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

The average predicted value of the outcome variable (v88) is 0.6, which means that the average prediction for v88 is 60%, indicating that the model predicts a Democratic vote share of around 60% on average across all districts.

(b)

Fit the same sort of model using the `brms` package with a t distribution, using the `brm` function with the student family. Again assess model fit.

```r
# Fit the model using the brm function and a Student's t-distribution
model_brms <- brm(v88 ~ v86 + inc88,
```

```
                  data = congress_data,
                  family = student(),  # Using the t-distribution
                  chains = 4,          # Number of MCMC chains
                  iter = 2000,         # Number of iterations per chain
                  warmup = 1000,       # Number of warm-up iterations
                  cores = 6)
```

```
## Compiling Stan program...
```

```
## Start sampling
```

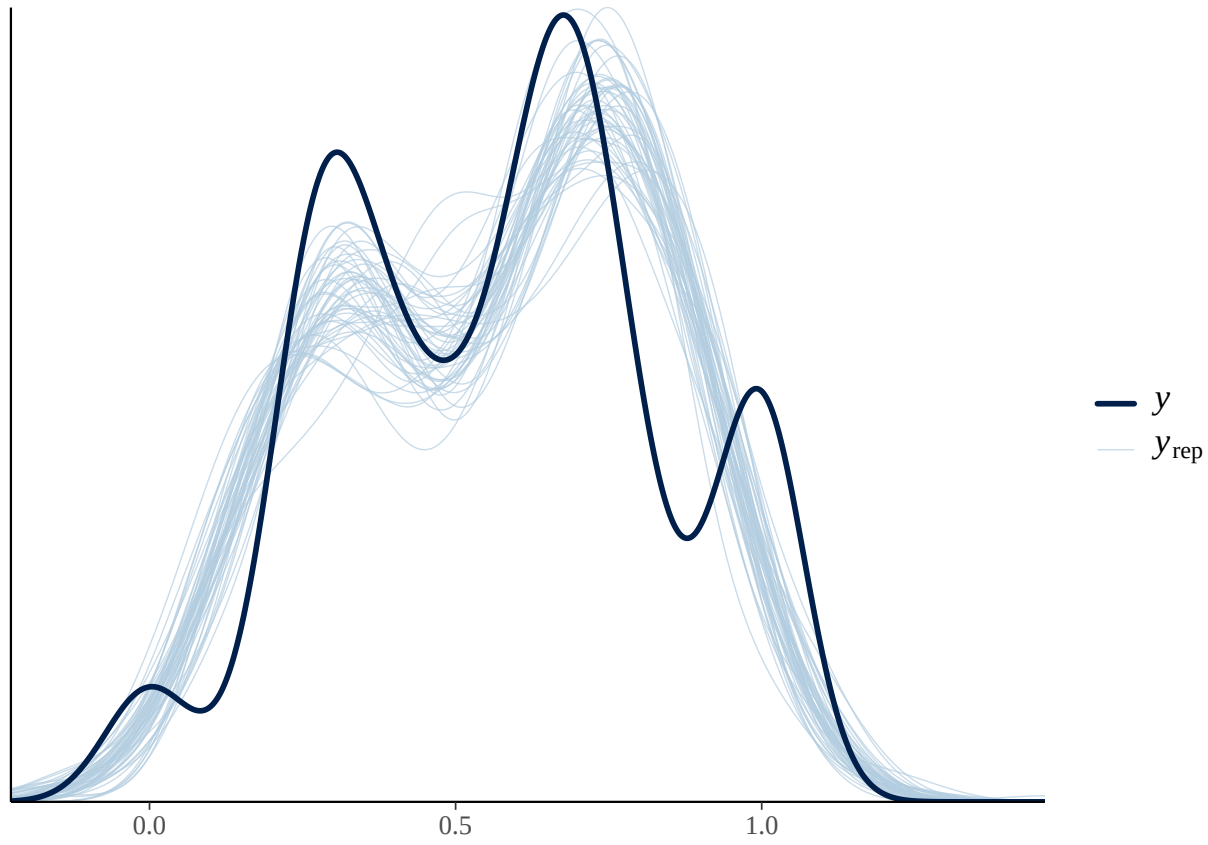
```
summary(model_brms)
```

```
##  Family: student
##   Links: mu = identity; sigma = identity; nu = identity
## Formula: v88 ~ v86 + inc88
##    Data: congress_data (Number of observations: 435)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Regression Coefficients:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.21      0.03     0.16     0.27 1.00     1691     1953
## v86           0.58      0.05     0.49     0.68 1.00     1630     1944
## inc88         0.10      0.01     0.07     0.12 1.00     1680     2043
##
## Further Distributional Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.09      0.01     0.08     0.11 1.00     1755     1933
## nu        3.45      1.07     2.10     5.94 1.00     1716     1768
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

(c)

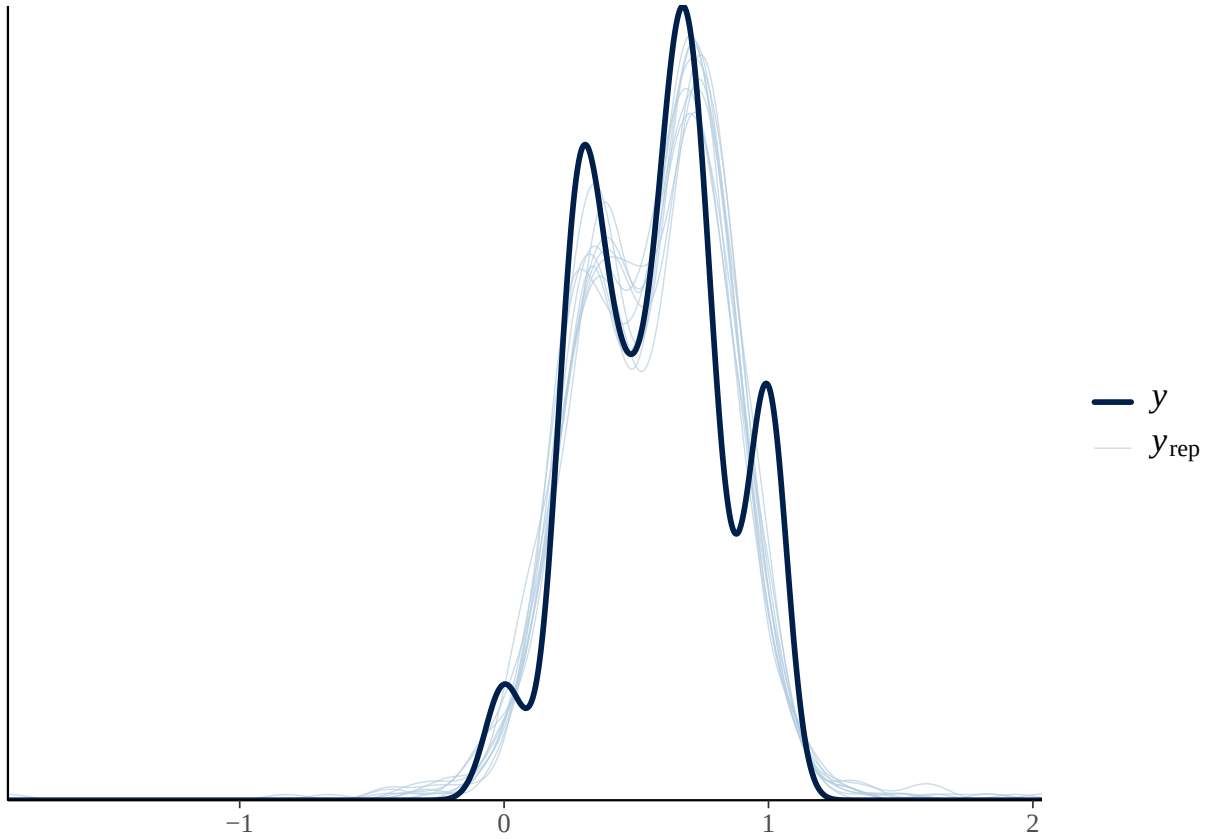
Which model do you prefer?

```
pp_check(model_stan)
```




```
pp_check(model_brms)
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```



Second Plot appears to represent a better fit overall because the predicted values (y_rep) align more closely with the observed data (y) across the main distribution.

15.9 Robust regression for binary data using the robit model

Use the same data as the previous example with the goal instead of predicting for each district whether it was won by the Democratic or Republican candidate.

(a)

Fit a standard logistic or probit regression and assess model fit.

(b)

Fit a robit regression and assess model fit.

(c)

Which model do you prefer?

15.14 Model checking for count data

The folder `RiskyBehavior` contains data from a study of behavior of couples at risk for HIV; see Exercise 15.1.

(a)

Fit a Poisson regression predicting number of unprotected sex acts from baseline HIV status. Perform predictive simulation to generate 1000 datasets and record the percentage of observations that are equal to 0 and the percentage that are greater than 10 (the third quartile in the observed data) for each. Compare these to the observed value in the original data.

(b)

Repeat (a) using a negative binomial (overdispersed Poisson) regression.

(c)

Repeat (b), also including ethnicity and baseline number of unprotected sex acts as inputs.

15.15 Summarizing inferences and predictions using simulation

Exercise 15.7 used a Tobit model to fit a regression with an outcome that had mixed discrete and continuous data. In this exercise you will revisit these data and build a two-step model: (1) logistic regression for zero earnings versus positive earnings, and (2) linear regression for level of earnings given earnings are positive. Compare predictions that result from each of these models with each other.