Stat 539 Homework 6

Kenny Flagg

April 4, 2017

1. *In murder trials in 20 Florida counties during 1976 and 1977, the death penalty was given in 19 out of 151 cases in which a white killed a white, in 0 out of 9 cases in which a white killed a black, in 11 out of 63 cases in which a black killed a white, and in 6 out of 103 cases in which a black killed a black (M. Radelet,* Am. Sociol. Rev.*, 46: 918-927, 1981). We will model these data using both log-linear models for contingency tables and logistic regression.*

   (a) *Enter these data into R as a* `data.frame` *with four columns:* `Defendant` *(race of defendant with levels "B" and "W"),* `Victim` *(race of victim with levels "B" and "W"),* `Penalty` *(whether or not death penalty was invoked with levels "N" and "Y"), and* `Freq` *(cell frequency). Your* `data.frame` *should have eight rows (not including variable names). Display this* `data.frame` *in R.*

```
> prob1 <- data.frame(
+   Defendant = factor(c('W', 'W', 'W', 'W', 'B', 'B', 'B', 'B'), levels = c('B', 'W')),
+   Victim = factor(c('W', 'W', 'B', 'B', 'W', 'W', 'B', 'B'), levels = c('B', 'W')),
+   Penalty = factor(c('Y', 'N', 'Y', 'N', 'Y', 'N', 'Y', 'N'), levels = c('N', 'Y')),
+   Freq = c(19, 132, 0, 9, 11, 52, 6, 97)
+ )
>
> print(prob1)

  Defendant Victim Penalty Freq
1         W      W       Y   19
2         W      W       N  132
3         W      B       Y    0
4         W      B       N    9
5         B      W       Y   11
6         B      W       N   52
7         B      B       Y    6
8         B      B       N   97
```

(b) *Use the R function* `xtabs` *with the* `data.frame` *format of the data to construct the partial tables needed to study the conditional association between defendant's race and the death penalty verdict, conditional on victim's race. Find and interpret the sample conditional odds ratios, adding 0.5 to each cell to reduce the impact of the zero cell count.*

```
> print(xtabs(Freq ~ Defendant + Penalty + Victim, data = prob1))

, , Victim = B

         Penalty
Defendant   N   Y
        B  97   6
        W   9   0

, , Victim = W

         Penalty
Defendant   N   Y
        B  52  11
        W 132  19
```

For black victims, $\widehat{OR}_B = \dfrac{(0+0.5)(97+0.5)}{(9+0.5)(6+0.5)} = 0.789$.

The odds of receiving the death penalty for a white defendant accused of killing a black victim are an estimated 21.1% lower than the odds of receiving the death penalty for a black defendant accused of killing a black victim.

For white victims, $\widehat{OR}_W = \dfrac{(19+0.5)(52+0.5)}{(132+0.5)(11+0.5)} = 0.672$.

The odds of receiving the death penalty for a white defendant accused of killing a white victim are an estimated 32.8% lower than the odds of receiving the death penalty for a black defendant accused of killing a white victim.

(c) *Find and interpret the sample marginal odds ratio between defendant's race and the death penalty verdict. Do these data exhibit Simpson's paradox? Explain.*

```
> print(xtabs(Freq ~ Defendant + Penalty, data = prob1))

         Penalty
Defendant   N   Y
        B 149  17
        W 141  19
```

For victims of either race, $\widehat{OR} = \dfrac{19 \times 149}{141 \times 17} = 1.181$.

The odds of receiving the death penalty for a white defendant accused of killing a black or white victim are an estimated 18.1% higher than the odds of receiving the death penalty for a black defendant accused of killing a black or white victim. The marginal odds of the death penalty are higher for whites than for blacks, but the conditional odds are lower for whites than for blacks regardless of the victim's race, so these data exhibit Simpson's paradox. This is because there were fewer cases with black victims than white

victims, the proportion of death sentences in cases with black victims was lower than for cases with white victims, and the proportion of white defendents was lower for cases with black victims than for cases with white victims.

(d) *Fit a logistic regression model which allows you to study the marginal association between defendant's race and the death penalty verdict. Use your fitted logistic regression model to answer the following questions.*

```
> # Use tidyr and dplyr to put the data in wide format.
> prob1_wide <- prob1 %>% spread(Penalty, Freq)
>
> prob1_marginal <- glm(cbind(Y, N) ~ Defendant, family = binomial,
+                       data = prob1_wide)
> print(xtable(summary(prob1_marginal)$coefficients, digits = 4))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.1707	0.2560	-8.4796	0.0000
DefendantW	0.1664	0.3539	0.4702	0.6382

i. *What is the estimated marginal odds ratio between defendant's race and the death penalty verdict? Show your work. (Hint: You should get the same answer as in part (c).)*

For cases with a victim of either race, the estimated odds ratio between the defendant's race and the death penalty verdict is

$$\widehat{OR} = \exp(0.1664) = 1.181$$

as seen in part (c).

ii. *Calculate and interpret an approximate 95% confidence interval for the marginal odds ratio of the death penalty comparing white defendants to black defendants.*

$$\exp(0.1664 \pm 1.96 \times 0.3539) = (0.590, 2.363)$$

We are 95% confident that the true odds of the death penalty for a white defendant accused of killing a black or white victim are between 41.0% lower and 136.3% higher than the odds of the death penalty for a black defendant accused of killing a black or white victim.

iii. *Is there significant statistical evidence to suggest that the defendant's race has an effect on the odds of the death penalty? Justify your answer.*

We are testing the null hypothesis of no association between the defendant's race and the odds of the death penalty ($\beta = 0$) against the alternative hypothesis that there is an association between the defendant's race and the odds of the death penalty ($\beta \neq 0$). There is no evidence of an association between the defendant's race and the odds of the death penalty ($z = 0.47$, p-value $= 0.64$).

(e) *Fit a logistic regression model which allows you to study the conditional association between defendant's race and the death penalty verdict (conditional on victim's race). Assume homogeneous association between defendant's race and the death penalty, conditional on victim's race.* Use your fitted logistic regression model *to answer the following questions.*

```
> prob1_conditional <- glm(cbind(Y, N) ~ Defendant + Victim, family = binomial,
+                          data = prob1_wide)
> print(xtable(summary(prob1_conditional)$coefficients, digits = 4))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.8421	0.4203	-6.7615	0.0000
DefendantW	-0.4402	0.4009	-1.0981	0.2722
VictimW	1.3242	0.5193	2.5498	0.0108

i. *What is the estimated conditional odds ratio between defendant's race and the death penalty for black victims? for white victims? Show your work. (Note that these answers will* not *match the answers in part (b).)*

For cases with black victims, the estimated odds ratio between the defendant's race and the death penalty verdict is

$$\widehat{OR} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1) = \exp(-0.4402) = 0.644.$$

For cases with white victims, the estimated odds ratio between the defendant's race and the death penalty verdict is

$$\widehat{OR} = \frac{\exp(\beta_0 + \beta_1 + \beta_2)}{\exp(\beta_0 + \beta_2)} = \exp(\beta_1) = \exp(-0.4402) = 0.644.$$

ii. *Calculate and interpret an approximate 95% confidence interval for the conditional odds ratio of the death penalty comparing white defendants to black defendants conditioned on victim's race.*

$$\exp(-0.4402 \pm 1.96 \times 0.4009) = (0.293, 1.413)$$

We are 95% confident that the true odds of the death penalty for a white defendant accused of killing a victim of a given race are between 70.7% lower and 41.3% higher than the odds of the death penalty for a black defendant accused of killing a victim of the same race.

iii. *Is there significant statistical evidence to suggest that the defendant's race has an effect on the odds of the death penalty after we control for the victim's race? Justify your answer.*

Here, we are testing the null hypothesis of no association between the defendant's race and the odds of the death penalty after accounting for the victim's race ($\beta = 0$) against the alternative hypothesis that there is an association between the defendant's race and the odds of the death penalty after accounting for the victim's race ($\beta \neq 0$). There is no evidence of an association between the defendant's race and the odds of the death penalty after accounting for the victim's race ($z = -1.10$, p-value $= 0.27$).

(f) *Fit the following four loglinear models (using the same notation as in the book), where D = Defendant, V = Victim, and P = Penalty: (D, V, P), (DV, VP), (DV, VP, DP), (DVP). You may use either R function* `glm` *or* `loglm`*. For each of the four fitted models, report the fitted conditional odds ratio(s) between Defendant and Penalty, conditioned on Victim.*

```
> `prob1_(D,V,P)` <- glm(Freq ~ Defendant + Victim + Penalty,
+                        family = poisson, data = prob1)
> `prob1_(DV,VP)` <- glm(Freq ~ Defendant * Victim + Victim * Penalty,
+                        family = poisson, data = prob1)
> `prob1_(DV,VP,DP)` <- glm(Freq ~ (Defendant + Victim + Penalty)^2,
+                        family = poisson, data = prob1)
> `prob1_(DVP)` <- glm(Freq ~ Defendant * Victim * Penalty,
+                      family = poisson, data = prob1)
```

For (D, V, P), $\widehat{OR} = 1$ because there is no DP term in the model.

For (DV, VP), $\widehat{OR} = 1$ because there is no DP term in the model.

For (DV, VP, DP), $\widehat{OR} = \exp\left(\hat{\gamma}_{WY}^{DP}\right) = \exp(-0.440) = 0.644$.

For (DVP), $\widehat{OR}_B = \exp\left(\hat{\gamma}_{WY}^{DP}\right) = \exp(-21.717) \approx 0$ for cases with black victims, and $\widehat{OR}_W = \exp\left(\hat{\gamma}_{WY}^{DP} + \hat{\delta}_{WWY}^{DVP}\right) = \exp(-21.717 + 21.332) = 0.680$ for cases with white victims.

(g) *Assess the goodness of fit for each of the four models fit in part (f). Which model would you choose and why?*

Model	G^2	DF	P-value
(D, V, P)	137.9294	4	<0.0001
(DV, VP)	1.8819	2	0.3903
(DV, VP, DP)	0.7007	1	0.4025
(DVP)	0.0000	0	–

There is very strong evidence that (D, V, P) is a poor fit. There is no evidence that (DV, VP) or (DV, VP, DP) are inadequate. (DVP) is the saturated model, so a better fit in not possible. I would use (DV, VP) because it is the simplest model that describes the data well.

(h) *For your chosen model from part (g), calculate an approximate 95% confidence interval for the conditional odds ratio(s) between Victim and Penalty, conditioned on Defendant. Show your work. How does this interval compare to the interval calculated in part (e)ii?*

```
> # It really bugs me how there's no easy way to have xtable pass p-values
> # through format.pval.
> print(xtable(summary(`prob1_(DV,VP)`)$coefficients, digits = 4))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.5797	0.1011	45.3141	0.0000
DefendantW	-2.4375	0.3476	-7.0126	0.0000
VictimW	-0.5876	0.1639	-3.5859	0.0003
PenaltyY	-2.8717	0.4196	-6.8431	0.0000
DefendantW:VictimW	3.3116	0.3786	8.7478	0.0000
VictimW:PenaltyY	1.0579	0.4635	2.2823	0.0225

For black defendents, the 95% confidence interval is

$$\exp(1.0579 \pm 1.96 \times 0.4635) = (1.161, 7.145).$$

For white defendents, the standard error is

```
> sqrt(c(0, 0, 0, 0, 1, 1) %*% vcov(`prob1_(DV,VP)`) %*% c(0, 0, 0, 0, 1, 1))

          [,1]
[1,] 0.5984848
```

so the 95% confidence interval is

$$\exp(3.3116 + 1.0579 \pm 1.96 \times 0.5985) = (0.891, 9.309).$$

These confidence intervals differ from the interval in (e)ii in that these intervals compare the odds of the death penalty between black and white *victims*, while the interval in (e)ii compares the odds of the death penalty between black and white *defendants*.

2. *Assume that λ has a Gamma distribution with mean μ and shape parameter $k > 0$. That is, the pdf of λ is:*

$$f(\lambda) = \frac{(k/\mu)^k}{\Gamma(k)} \lambda^{k-1} e^{-k\lambda/\mu}$$

and that $Y|\lambda \sim Pois(\lambda)$. Show that the marginal probability mass function of Y (given μ and k)

(a) *is equal to the negative binomial pmf*

$$p(y|\mu, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k, \qquad y = 0, 1, 2, \ldots$$

The joint distribution function of Y and λ is

$$\begin{aligned} p(y|\lambda) f(\lambda) &= e^{-\lambda} \frac{\lambda^y}{y!} \frac{(k/\mu)^k}{\Gamma(k)} \lambda^{k-1} e^{-k\lambda/\mu} \\ &= \frac{(k/\mu)^k}{\Gamma(k)\Gamma(y+1)} \lambda^{y+k-1} e^{-\lambda\left(\frac{\mu+k}{\mu}\right)} \end{aligned}$$

so the marginal distribution of Y is

$$\begin{aligned} p(y|\mu, k) &= \int_0^\infty p(y|\lambda) f(\lambda) d\lambda \\ &= \int_0^\infty \frac{(k/\mu)^k}{\Gamma(k)\Gamma(y+1)} \lambda^{y+k-1} e^{-\lambda\left(\frac{\mu+k}{\mu}\right)} d\lambda \\ &= \frac{(k/\mu)^k}{\Gamma(k)\Gamma(y+1)} \Gamma(y+k) \left(\frac{\mu}{\mu+k}\right)^{y+k} \\ &\quad \times \int_0^\infty \frac{1}{\Gamma(y+k)} \left(\frac{\mu+k}{\mu}\right)^{y+k} \lambda^{y+k-1} e^{-\lambda\left(\frac{\mu+k}{\mu}\right)} d\lambda \qquad (1) \\ &= \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{\mu}{\mu+k}\right)^k \left(\frac{k}{\mu}\right)^k \\ &= \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k, \qquad y = 0, 1, 2, \ldots \end{aligned}$$

because the expression inside the integral in (**??**) is a Gamma density. The result is a negative binomial mass function.

(b) *and is a member of the exponential dispersion family.*

The probability mass function can be written as

$$\begin{aligned} p(y|\mu,k) &= \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{\mu}{\mu+k}\right)^y\left(\frac{k}{\mu+k}\right)^k \\ &= \exp\left[y\log\left(\frac{\mu}{\mu+k}\right)+k\log\left(\frac{k}{\mu+k}\right)+\log\left(\frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\right)\right] \\ &= \exp\left[y\log\left(\frac{\mu}{\mu+k}\right)+k\log\left(1-\frac{\mu}{\mu+k}\right)+\log\left(\frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\right)\right] \\ &= \exp\left[y\log\left(\frac{\mu}{\mu+k}\right)+\log\left(1-e^{\log\left(\frac{\mu}{\mu+k}\right)}\right)/\left(\frac{1}{k}\right)+\log\left(\frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\right)\right] \end{aligned}$$

which is the exponential family form in the parameterization of Jørgensen (1987), with $\theta = \log\left(\frac{\mu}{\mu+k}\right)$, $b(\theta) = -\log(1-e^\theta)$, $\phi = k$, $a(\phi) = \frac{1}{\phi}$, and $c(y,\phi) = \log\left(\frac{\Gamma(y+\phi)}{\Gamma(\phi)\Gamma(y+1)}\right)$.

3. *In this problem we will use data obtained from a sexually transmitted disease (STD) clinic to determine factors that are associated with the rate of reinfection of STDs. The dataset contains aggregated (grouped) data including the total number of reinfections per group, the total amount of time each group was followed for, and various other covariates. The data can be found on the course website at* `http://www.math.montana.edu/shancock/courses/stat539/data/stdgrp.txt`*. The following variables are available in the dataset:*

Variable	Description
`white`	Indicator of white race (1=white, 0=black)
`married`	Marital status (1=single, 2=married, 3=divorced/separated)
`agegrp`	Age group (categorized as [13,19], (19,22], and (22,48])
`edugrp`	Years of schooling (categorized as [6,11.9], (11.9,12.9], (12.9,18])
`inftype`	Initial infection (1= gonorrhea, 2=chlamydia, 3=both)
`npartnr`	Number of sexual partners (categorized as [0,1], (1,2], (2,3], and (3,19])
`condom`	Condom use (1=always, 0=sometimes/never)
`n.reinfect`	Total number of reinfections observed in the group
`yrsfu`	Total number of years individuals in the group were followed for

(a) *If we model the total number of reinfections observed in a group (`n.reinfect`) as a Poisson random variable, what variable in the data set should serve as the offset term?*

The offset should be log(`yrsfu`) because, conditional on the other predictors and assuming a homogeneous process, the mean number of reinfections observed would be proportional to the time (in person-years) that the individuals in the group were followed.

(b) *Write down the Poisson regression model that shows how the rate of reinfection varies with race, education group, initial infection and condom use (without interaction terms). Explain, using your model, how the rate is related to the actual count of reinfections.*

For race i, education group j, initial infection type k, and condom use status l, the number of reinfections y_{ijkl} is modeled as

$$y_{ijkl} \sim \text{Poisson}(\mu_{ijkl});$$
$$\log(\lambda_{ijkl}) = \log(t_{ijkl}) + \beta_0 + \beta_i^R + \beta_j^E + \beta_k^C + \beta_l^I$$

where λ_{ijkl} is the expected number of reinfection per person-year, t_{ijkl} is the number of person-years the group was observed, and $\mu_{ijkl} = \lambda_{ijkl} t_{ijkl}$ is the expected number of reinfections in t_{ijkl} person-years.

(c) *Fit the model in part (b) (show R code and output) and choose two estimated coefficients to interpret in context of the problem.*

```
> prob3 <- read.table('http://www.math.montana.edu/shancock/courses/stat539/data/stdgrp.txt',
+                     header = TRUE)
>
> prob3_rate <- glm(n.reinfect ~ white + edugrp + factor(inftype) + condom,
+                   offset = log(yrsfu), family = poisson, data = prob3)
> print(xtable(summary(prob3_rate)$coefficients, digits = 4))
```

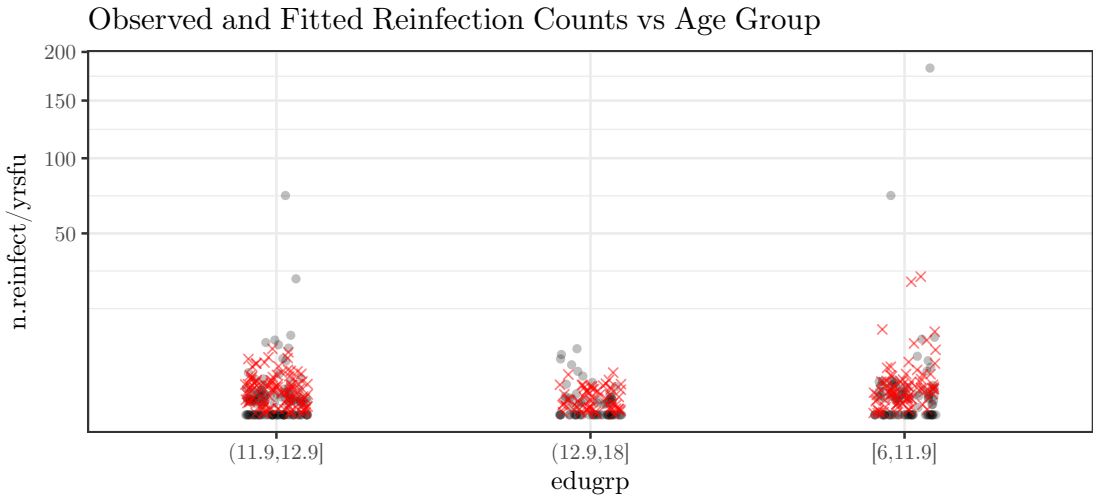
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.5020	0.1465	-3.4272	0.0006
white	-0.2935	0.1296	-2.2641	0.0236
edugrp(12.9,18]	-0.3450	0.1922	-1.7950	0.0727
edugrp[6,11.9]	0.2226	0.1165	1.9103	0.0561
factor(inftype)2	-0.3747	0.1478	-2.5351	0.0112
factor(inftype)3	-0.2968	0.1452	-2.0444	0.0409
condom	-0.3724	0.1150	-3.2376	0.0012

For people of a given age group, infection type, and condom use, the annual reinfection rate for whites is estimated to be $1 - \exp(-0.2935) = 25.4\%$ lower for whites than for blacks.

For people of a given race, age group, and infection type, the annual reinfection rate is estimated to be $1 - \exp(-0.3724) = 31.1\%$ lower for people who always use condoms than for people who sometimes or never use condoms.

(d) *Create an appropriate plot to examine the data for overdispersion. Refit your model accounting for overdispersion using a quasi-Poisson family. How do your conclusions change (if at all) after accounting for overdispersion?*

```
> ggplot(prob3, aes(x = edugrp)) +
+   geom_jitter(aes(y = n.reinfect / yrsfu),
+               height = 0, width = 0.1, shape = 16, col = '#00000040') +
+   geom_jitter(aes(y = fitted(prob3_rate)),
+               height = 0, width = 0.1, shape = 4, col = '#ff000080') +
+   scale_y_sqrt() + # Square root scale to remove some of the skew.
+   ggtitle('Observed and Fitted Reinfection Counts vs Age Group')
```




```
> prob3_quasi <- glm(n.reinfect ~ white + edugrp + factor(inftype) + condom,
+                    offset = log(yrsfu), family = quasipoisson, data = prob3)
> print(xtable(summary(prob3_quasi)$coefficients, digits = 4,
+       caption = sprintf('Dispersion parameter = %.2f',
+                         summary(prob3_quasi)$dispersion)))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.5020	0.3330	-1.5074	0.1330
white	-0.2935	0.2947	-0.9958	0.3203
edugrp(12.9,18]	-0.3450	0.4370	-0.7895	0.4306
edugrp[6,11.9]	0.2226	0.2649	0.8402	0.4016
factor(inftype)2	-0.3747	0.3360	-1.1150	0.2659
factor(inftype)3	-0.2968	0.3301	-0.8992	0.3694
condom	-0.3724	0.2615	-1.4240	0.1557

Table 1: Dispersion parameter = 5.17

On the plot, the circles represent the observed rates and the crosses represent the fitted rates. More large rates and rates of zero were observed than the fitted model predicts, and the dispersion paramter is large, indicating the presence of overdispersion. After correcting the standard errors for overdispersion, none of the predictors appear to be associated with the reinfection rate.

(e) *Do the grouped reinfection counts appear to follow a Poisson distribution? Why or why not?*

The plot in (d) shows more observed rates of zero than should be expected for a Poisson distribution. The counts do not appear to follow a Poisson distribution and might be better described by a zero-inflated model.

(f) *If we control for race and education, does condom use have a significant effect on the rate of re-infection? Conduct the appropriate likelihood ratio test (accounting for overdispersion) to address this question and write a conclusion of the test in context of the problem.*

```
> prob3_reduced <- glm(n.reinfect ~ white + edugrp + factor(inftype),
+                      offset = log(yrsfu), family = quasipoisson, data = prob3)
> print(xtable(anova(prob3_reduced, prob3_quasi, test = 'LRT'),
+              digits = c(0, 0, 2, 0, 4, 4)), include.rownames = FALSE)
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
252	335.84			
251	324.98	1	10.8545	0.1473

This is a test of the null hypothesis that condom use is not associated with the reinfection rate after accounting for race and education ($\beta_3 = 0$) versus the alternative hypothesis that condom use is associated with the reinfection rate after accounting for race and education ($\beta_3 \neq 0$). There is no evidence of an association between condom use and the reinfection rate after accounting for race and education ($\chi^2_1 = 10.85$, p-value $= 0.15$).