STA303H1 - Assignment 1

2023-07-23

Question 1

a)

No. First note that $E[y_i|x] = E[\beta_0 + \beta_1 x_i + \epsilon_i|x] = E[\beta_0 + \beta_1 x_i|x]$ since assumption 3 still holds.

To show $\hat{\beta}_1$ is still unbiased, define $k_i = \frac{x_i - \overline{x}}{SXX}$ $\forall i$, where $SXX = \sum_i (x_i - \overline{x})^2$. Notice that $\sum_i k_i = \frac{1}{SXX}(\sum_i x_i - n\overline{x}) = 0$. Then,

$$
\begin{aligned}
\hat{\beta}_1 &= \sum_i k_i(y_i - \overline{y}) = \sum_i k_i y_i - \overline{y}\sum_i k_i = \sum_i k_i y_i - 0 \\
E[\hat{\beta}_1|x] &= E[\sum_i k_i y_i|x] = \sum_i k_i E[y_i|x] = \sum_i k_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_i k_i + \beta_1 \sum_i k_i x_i \\
&= 0 + \beta_1(\sum_i k_i x_i - \overline{x}\sum_i k_i) = \beta_1 \sum_i k_i(x_i - \overline{x}) = \beta_1 \frac{\sum_i (x_i - \overline{x})^2}{SXX} \\
&= \beta_1
\end{aligned}
$$

as needed. To show $\hat{\beta}_0$ is still unbiased,

$$
\begin{aligned}
E[\hat{\beta}_0|x] &= E[\overline{y} - \hat{\beta}_1\overline{x}|x] = E[\overline{y}|x] - \overline{x}E[\hat{\beta}_1|x] = E[\frac{1}{n}\sum_i y_i|x] - \overline{x}\beta_1 = \frac{1}{n}\sum_i E[y_i|x] - \overline{x}\beta_1 \\
&= \frac{1}{n}\sum_i (\beta_0 + \beta_1 x_i) - \overline{x}\beta_1 = \frac{1}{n}n\beta_0 + \frac{1}{n}\beta_1 \sum_i x_i - \overline{x}\beta_1 = \beta_0 + \beta_1\overline{x} - \overline{x}\beta_1 \\
&= \beta_0
\end{aligned}
$$

as needed.

b)

The assumption of homoskedasticity is violated since the variance is not constant but rather varies depending on x_i.

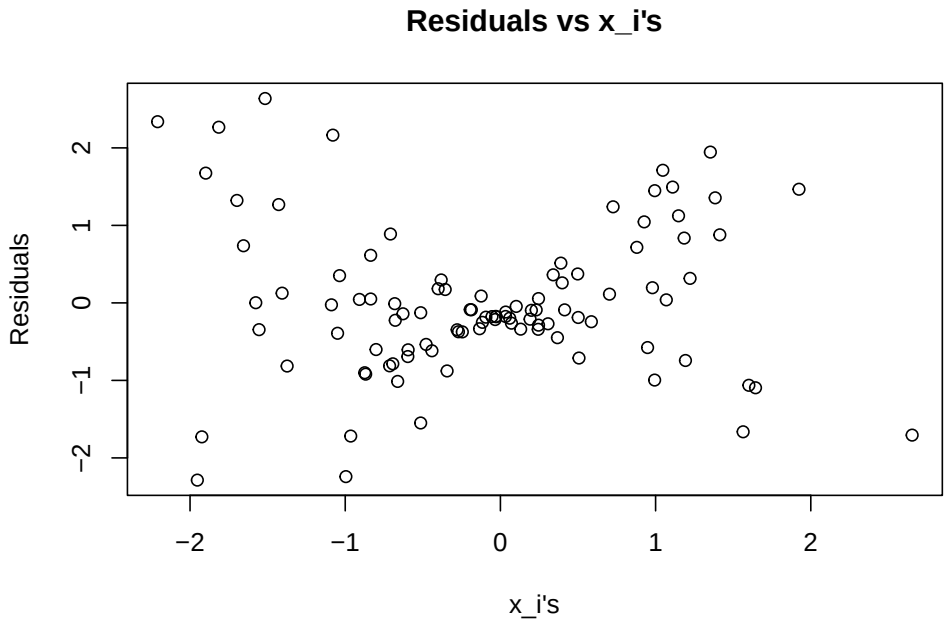
c)

```
set.seed(1007941426)
df_c <- data.frame(matrix(ncol=2, nrow=0))
colnames(df_c) <- c("x", "y")
x = rnorm(100, mean=0, sd=1)
```

```r
for (i in 1:100) {
  e_i = rnorm(1, mean=0, sd=abs(x[i]))
  y_i = 1 + 2*x[i] + e_i

  # Adds x_i and y_i to the dataframe
  df_c[nrow(df_c) + 1,] = c(x[i], y_i)
}
```

d)

```r
fit_d <- lm(y ~ x, data=df_c)
```



The residuals roughly follow the absolute value function for x values between -2.5 and 3, which agrees with how the random error terms are simulated. Since the residuals have a pattern, we know homoskedasticity is violated.

e)

```r
set.seed(1007941426)
beta1_est <- as.vector(NULL)
within_CI <- as.vector(NULL)
n <- 100

for (i in 1:1000){
  # Simulate data
  e <- rnorm(n, 0, abs(x))
  y <- 1 + 2*x + e

  # Fit model
  fit_e <- lm(y ~ x, data=data.frame(x, y))

  # Store the estimated coefficient for beta_1
  beta1_est[i] <- fit_e$coefficients["x"]
```

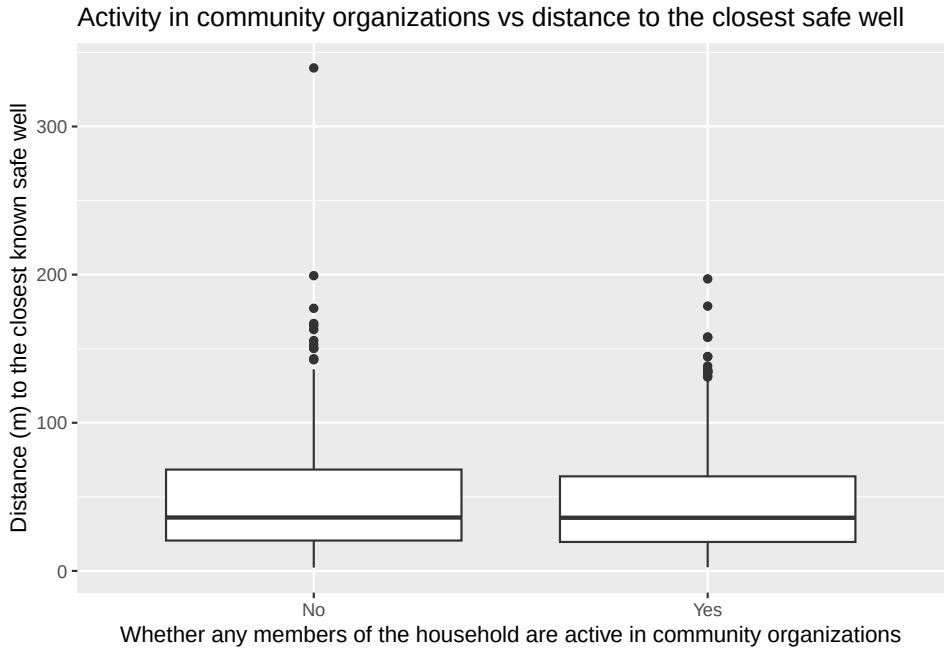
```
  ci <- confint(fit_e, parm='x', level=0.95)

  # Does the confidence interval contain the true value of beta_1?
  within_CI[i] <- 2 >= ci[1] & 2 <= ci[2]
}

mean_beta1 <- mean(beta1_est)
proportion_contains <- mean(within_CI)
```

The mean of the $\hat{\beta}_1$'s is 2.00414. The proportion of times my confidence intervals contain β_1 is 0.768.

The mean $\hat{\beta}_1$ is very close to the true $\beta_1 = 2$, meaning the $\hat{\beta}_1$'s are not biased, which is consistent with my answer in a). The confidence interval for β_1, however, is not a true 95% confidence interval since it contains $\beta_1 = 2$ only 76.8% of the time. This is also consistent with my answer in a) since heteroskedasticity means the estimate of the variance is inaccurate, which means the confidence interval is also inaccurate.

Question 2

a)



The boxplot shows that both groups have similar median distances to the closest safe well. The third and fourth quartile distances of the "No" group are slightly greater than those for the "Yes" group. Both groups have outliers, but the "No" group has an extreme outlier with a distance of well above 300 meters. Thus, the boxplot suggests that there may be a slight interaction between activity in community organizations and distance from safe wells.

b)

```
# Counts # of missing values in each column
colSums(is.na(arsenic))
```

```
##  switch arsenic    dist   assoc    educ
##       0      11       0       0       0
```

```
# Total # of observations
nrow(arsenic)
```

```
## [1] 500
```

Arsenic has 11 missing values out of 500 total, or 2.2% missing data.

c)

The mechanism is missing completely at random. The bumpiness of the roads to the lab is random and unrelated to the data, meaning whether a test tube shatters or not (equivalently, whether 'arsenic' is missing) depends on pure chance.

d)

```
arsenic <- na.omit(arsenic)
```

I perform complete case analysis by deleting rows with missing data. Since the missing data mechanism is MCAR, the estimates of means, regression coefficients, and variances are unbiased. In addition, the proportion of missing data (2.2%) is small, meaning the dataset is unlikely to be significantly affected.

a):

```
mod_a <- glm(switch ~ arsenic + dist + assoc + educ, family=binomial(link="logit"), data=arsenic)
summary(mod_a)
```

```
##
## Call:
## glm(formula = switch ~ arsenic + dist + assoc + educ, family = binomial(link = "logit"),
##     data = arsenic)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.25814    0.25376  -1.017  0.30904
## arsenic      0.53634    0.10916   4.913 8.96e-07 ***
## dist        -0.01059    0.00268  -3.951 7.78e-05 ***
## assocYes    -0.15359    0.19296  -0.796  0.42604
## educ         0.06278    0.02430   2.584  0.00977 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 666.35  on 488  degrees of freedom
## Residual deviance: 621.45  on 484  degrees of freedom
## AIC: 631.45
##
## Number of Fisher Scoring iterations: 4
```

```
# Calculates the mean of each numeric variable
colMeans(select_if(arsenic, is.numeric))
```

```
##     switch    arsenic       dist       educ
##  0.5766871  1.6569734 47.9779468  4.7852761
```

Note: in all cases, we hold other covariates constant when interpreting the coefficient of a variable.

The probability of switching wells when the current well has no arsenic, the distance to the nearest safe well is 0 meters, the household is not involved in community associations, and the household head has no education is 0.436.

A 1 unit increase in the arsenic level of a well near the mean level is associated with a 11.0% increased probability of the household switching wells.

A 1 unit increase in the distance to the closest safe well near the mean distance is associated with a 0.229% decreased probability of the household switching wells.

A household active in community organizations is 3.73% less likely to switch wells than a household that is not active.

A 1 unit increase in the education level of the household head near the mean level is associated with a 1.57% increased probability of the household switching wells.

b):

```
mod_b <- glm(switch ~ arsenic + dist + assoc + educ + dist:assoc, family=binomial(link="logit"), data=a
summary(mod_b)
```

```
##
## Call:
## glm(formula = switch ~ arsenic + dist + assoc + educ + dist:assoc,
##     family = binomial(link = "logit"), data = arsenic)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.396427   0.276154  -1.436  0.15114
## arsenic        0.536930   0.109271   4.914 8.94e-07 ***
## dist          -0.007903   0.003351  -2.358  0.01835 *
## assocYes       0.156686   0.313318   0.500  0.61701
## educ           0.063564   0.024315   2.614  0.00894 **
## dist:assocYes -0.006558   0.005233  -1.253  0.21014
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 666.35  on 488  degrees of freedom
## Residual deviance: 619.86  on 483  degrees of freedom
## AIC: 631.86
##
## Number of Fisher Scoring iterations: 4
```

The probability of switching wells when the current well has no arsenic, the distance to the nearest safe well is 0 meters, the household is not involved in community associations, and the household head has no education is 0.402.

A 1 unit increase in the arsenic level of a well near the mean level is associated with a 11.6% increased probability of the household switching wells.

Provided that the household is not active in community organizations, a 1 unit increase in the distance to the closest safe well near the mean distance is associated with a 0.170% decreased probability of the household switching wells.

A household active in community organizations is 3.82% more likely to switch wells than a household that is not active.

A 1 unit increase in the education level of the household head near the mean level is associated with a 1.59% increased probability of the household switching wells.

Provided that the household is active in community organizations, a 1 unit increase in the distance to the closest safe well near the mean distance is associated with a 3.60% decreased probability of the household switching wells.

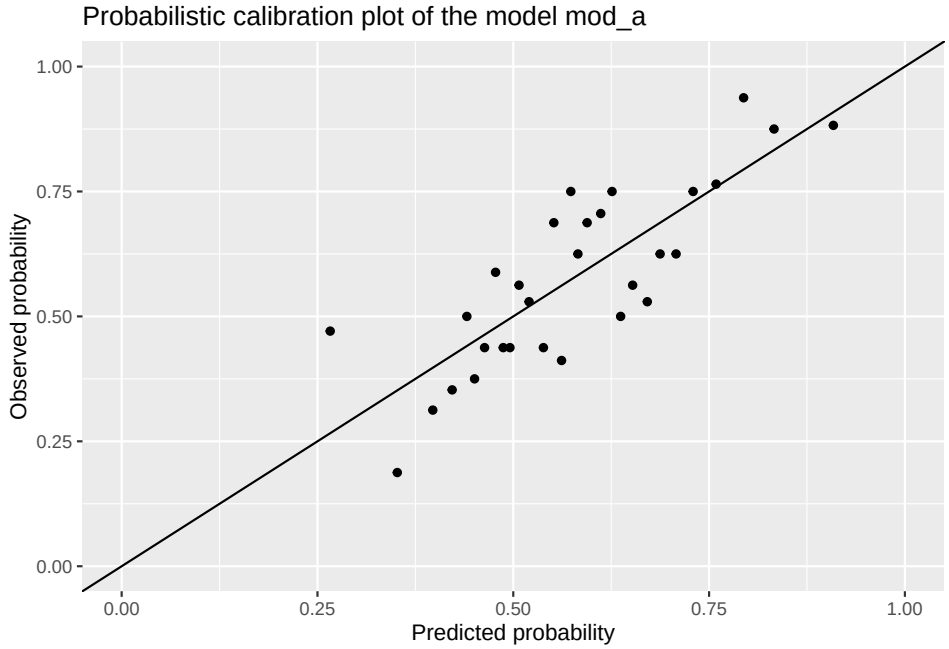
e)

```
lrtest(mod_a, mod_b)
```

```
## Likelihood ratio test
##
## Model 1: switch ~ arsenic + dist + assoc + educ
## Model 2: switch ~ arsenic + dist + assoc + educ + dist:assoc
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -310.73
## 2   6 -309.93  1 1.5883     0.2076
```

I choose mod_a (Model 1) with only additive predictors since by the LRT, there is no statistically significant change in the log-likelihood if the interaction term between 'dist' and 'assoc' is removed. Thus, we fail to reject the null hypothesis and conclude that the interaction term is not needed in the model.

f)



The calibration plot of mod_a with 30 bins shows that overall, the average fitted probabilities are somewhat different from the actual observed probabilities. Compared to the actual observed probabilities, about half of the bins has noticeably smaller average fitted probabilities, and about half has noticeably larger average fitted probabilities. There are only around 5 bins that lie on or near the line, meaning they have similar fitted and observed probabilities. Thus, the model is a moderate fit to the data.

```
# Brier score
mean((arsenic$switch - predict(mod_a, type="response"))^2)
```

```
## [1] 0.2230314
```

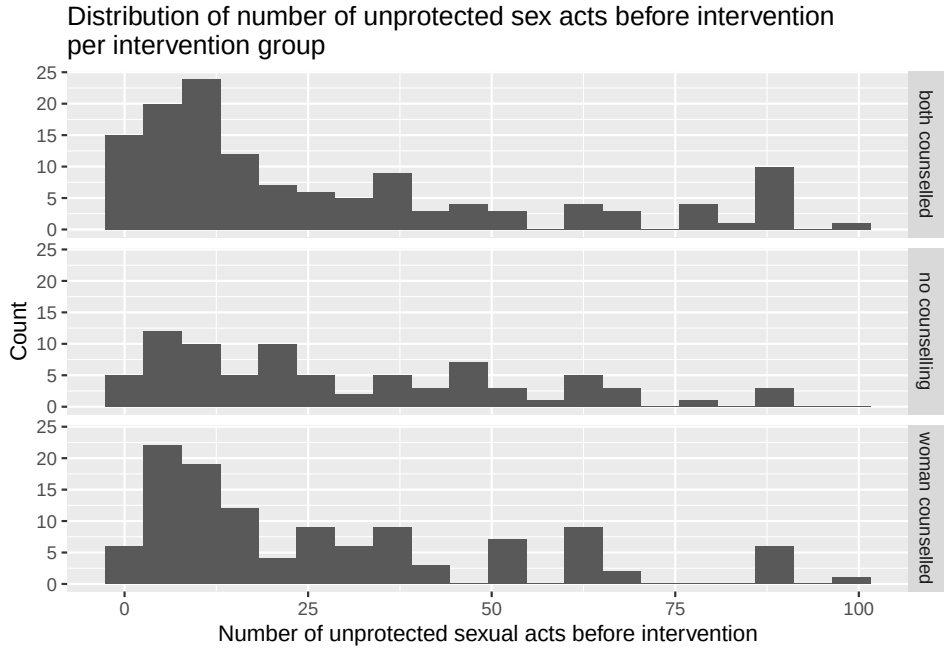
The Brier score, a measure of probabilistic accuracy, is 0.223. This is close to 0.25, which indicates a random guess (i.e., the predicted probability is 0.5). Thus, the model is a moderate fit to the data, since its performance is not substantially better than a model with random guesses.

Question 3

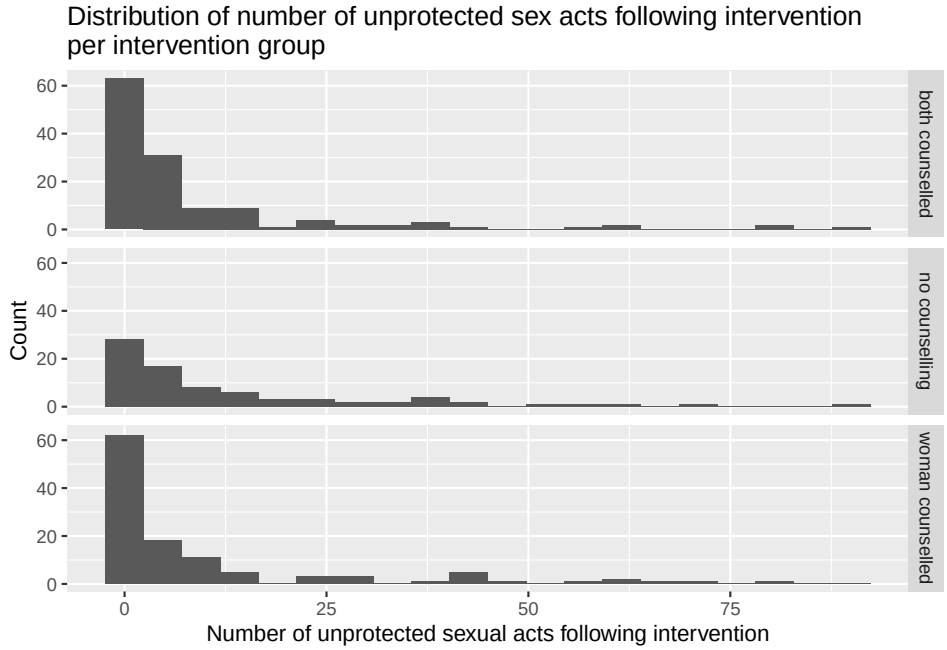
a)

```
hiv <- hiv[hiv$fupacts < 200 & hiv$bupacts < 200,]
```

I removed observations with a 'bupacts' or 'fupacts' value of ≥ 200, which are outliers since all the other observations have values between 0 and 100.



Before intervention, the distribution of the number of unprotected sex acts is spread out between 0 to 100 for all intervention groups. There is a greater concentration of values near approximately 10 in the 'both counseled' and 'woman counseled' groups, while the distribution of values for the 'no counseling' group is more uniform. There appears to be lower overall counts for 'no counseling' compared to the other two groups.



After intervention, the distribution of the number of unprotected sex acts is noticeably more concentrated near 0 for all groups compared to before intervention. In particular, the mode number of sex acts for all three groups is near 0. Thus, the exploratory data analysis suggests that intervention may be effective on reducing an individual's number of unprotected sexual acts.

b)

A Binomial model is used when the outcome is a binary categorical variable. Since a count variable is not binary but rather discrete quantitative, a Poisson model is preferred for modeling counts.

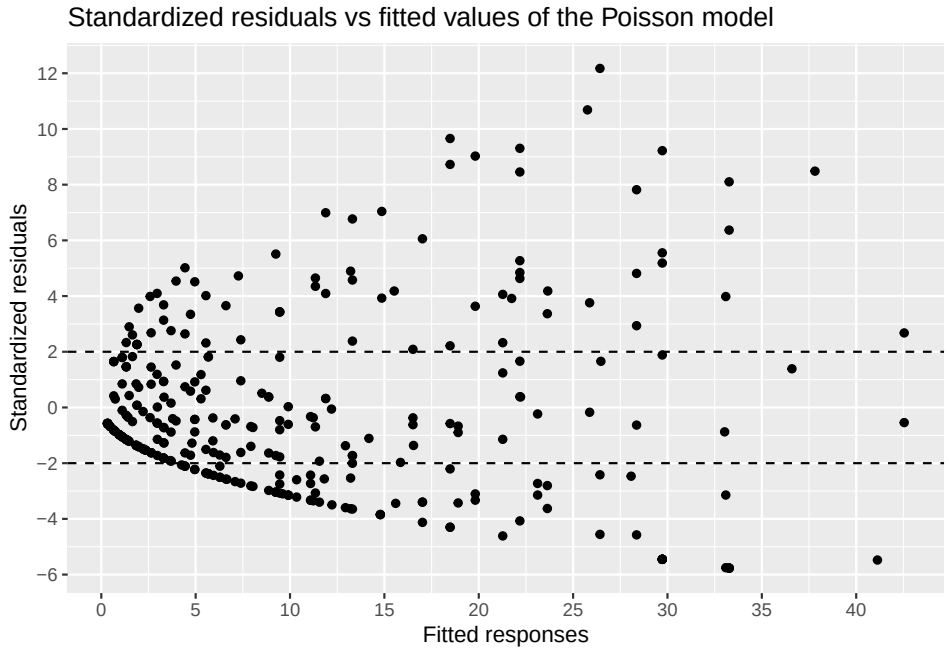
c)

```
mod_c <- glm(fupacts ~ intervention, data=hiv, offset=log(bupacts), family=poisson(link="log"))
summary(mod_c)
```

```
##
## Call:
## glm(formula = fupacts ~ intervention, family = poisson(link = "log"),
##     data = hiv, offset = log(bupacts))
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.10779    0.02895 -38.264  < 2e-16 ***
## interventionno counselling    0.35850    0.04189   8.558  < 2e-16 ***
## interventionwoman counselled  0.11257    0.04135   2.722  0.00648 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3715.2  on 325  degrees of freedom
## Residual deviance: 3641.3  on 323  degrees of freedom
## AIC: 4487.1
##
## Number of Fisher Scoring iterations: 5
```

I chose my offset to be 'bupacts', or the number of unprotected sexual acts prior to intervention, since we are interested in the reduction of unprotected sex acts due to intervention. This reduction can be expressed as a proportion of 'fupacts' to 'bupacts', since the value of 'bupacts' varies and we want to be able to compare the reduction for all individuals.

d)



In the standardized residuals plot, many observations lie outside of the range $(-2, 2)$. This means that the observed variance of y_i is greater than $\hat{\mu_i}$ (and hence μ_i), instead of being equal to it as assumed by the Poisson model. This indicates overdispersion. In the plot, I also notice several 'curves' of points, which may be because 'fupacts' is a discrete numerical variable.

```
# Estimated dispersion
# sum((hiv$fupacts - mean(hiv$fupacts))^2 / mean(hiv$fupacts)) / (nrow(hiv) - 3) # Incorrect formula
sum((hiv$fupacts - mod_c$fitted.values)^2 / mod_c$fitted.values) / (nrow(hiv) - 3) # Actual formula
```

```
## [1] 10.00741
```

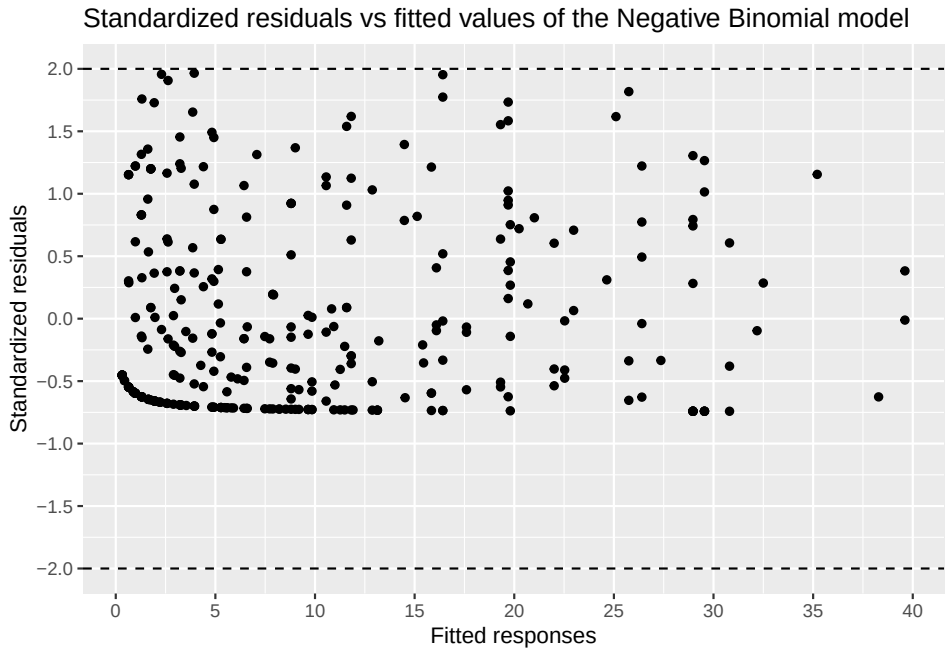
The estimated dispersion for the data is 29.3, which is greater than the theoretical dispersion parameter for a Poisson model of 1.

e)

```
mod_e <- glm.nb(fupacts ~ intervention + offset(log(bupacts)), data=hiv)
summary(mod_e)
```

```
##
## Call:
## glm.nb(formula = fupacts ~ intervention + offset(log(bupacts)),
##     data = hiv, init.theta = 0.5594161658, link = log)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -1.13361    0.12755  -8.888   <2e-16 ***
## interventionno counselling    0.31278    0.20307   1.540    0.123
## interventionwoman counselled  0.01974    0.18521   0.107    0.915
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.5594) family taken to be 1)
```

```
##
##     Null deviance: 353.98  on 325  degrees of freedom
## Residual deviance: 351.08  on 323  degrees of freedom
## AIC: 1842.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.5594
##           Std. Err.:  0.0563
##
##  2 x log-likelihood:  -1834.4080
```



All of the observations lie within the range $(-2, 2)$. This indicates that compared to a Poisson model, the Negative Binomial model is a much better fit. Similar to the standardized residuals plot of the Poisson model, there appears to be several 'curves' of points, and all standardized residuals are greater than or approximately equal to -0.75.

f)

All rates here are ratios with respect to the number of unprotected sexual acts before intervention.

The estimated rate of unprotected sexual acts following intervention is 0.322 when both partners are counseled.

The estimated relative change in the rate of unprotected sexual acts following intervention when neither partner is counseled compared to when both are counseled is 36.7%.

The estimated relative change in the rate of unprotected sexual acts following intervention when just the female is counseled compared to when both are counseled is 1.99%.

There is evidence that the intervention is effective. The intervention group where both partners were counseled has the lowest rate of unprotected sexual acts following intervention compared to before intervention, and this rate is < 1, indicating a reduction. As well, the rate when just the female is counseled is also lower than the rate when neither is counseled.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: May 5, 2021)