Homework 1

Joshua Ingram

8/30/2020

Problem 1

1.

```
fit1_chile <- multinom(vote ~ age + sex + region + statusquo, data = chile)
```

```
## # weights:  36 (24 variable)
## initial  value 3490.689201
## iter  10 value 2282.918477
## iter  20 value 2154.749143
## iter  30 value 2122.914761
## final  value 2122.850427
## converged
```

```
fit1_chile
```

```
## Call:
## multinom(formula = vote ~ age + sex + region + statusquo, data = chile)
##
## Coefficients:
##   (Intercept)         age       sexM   regionM     regionN    regionS
## N  -0.1295192 0.005397365  0.6993290 0.8411402 -0.30822679 0.33434661
## U   0.1589056 0.030373441 -0.2879216 1.3387552 -0.73715898 0.09402906
## Y  -0.3884494 0.025373911 -0.1039942 1.5077273  0.08062193 0.41435538
##     regionSA  statusquo
## N -0.0860732 -1.8230660
## U  0.0771308  0.3338119
## Y  0.2254778  1.8756710
##
## Residual Deviance: 4245.701
## AIC: 4293.701
```

"will abstain" is the baseline category.

2.

$$Y_i \sim_{ind.} Multinomial(p_{i,1}, p_{i,2}, p_{i,3}, p_{i,4})$$

Where 1 = "will vote no", 2 = "undecided", 3 = "will vote yes", 4 = "will abstain"

$$log(\frac{p_{i,j}}{p_{i,4}}) = \beta_{0,j} + \beta_{1,j} age_i + \beta_{2,j} sex_i + \beta_{3,j} region_i + \beta_{4,j} statusquo_i,\ j = 1, 2, 3$$

$$p_{i,4} = 1 - \sum_{j=1}^{3} p_{i,j}$$

3.

We use the Likelihood Ratio Test (LRT) to test each predictor "as a whole."

Example of hypothesis test for the variable age age (β_1):

$H_0 \ : \ \beta_{1,1} = \beta_{1,2} = \beta_{1,3} = 0$

$H_A \ : \ \{\exists\ \beta_{1,j} \neq 0 \mid j = 1, 2, 3\}$

```
Anova(fit1_chile)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: vote
##           LR Chisq Df Pr(>Chisq)
## age          41.55  3  4.998e-09 ***
## sex          59.22  3  8.627e-13 ***
## region       30.92 12   0.002028 **
## statusquo  1910.56  3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

All of our predictor variables are statistically significant according to the Likelihood Ratio Test.

4.

a.

$$log(\frac{\hat{p}_{yes}}{\hat{p}_{abstain}}) = -0.388 + 0.025 age_i - 0.104 I_{sexM,i} + 1.51 I_{regionM,i} + 0.08 I_{regionN,i} + 0.41 I_{regionS,i} + 0.23 I_{regionSA,i} + 1.88 stquo_i$$

where $I_{sexM} \in \{0 = female, 1 = male\}$ *and* $I_{regionK} \in \{0 = \text{not in region K}, 1 = \text{in region K}\},\ K = M, N, S, SA$

b.

```
z <- summary(fit1_chile)$coefficients/summary(fit1_chile)$standard.errors
p_val <- (1 - pnorm(abs(z), 0, 1)) * 2
round(p_val,8)
```

```
##   (Intercept)        age       sexM    regionM    regionN   regionS  regionSA
## N   0.6724581 0.39983795 0.00005595 0.28826706 0.28223894 0.1905066 0.7044145
## U   0.5890860 0.00000149 0.09686135 0.07858723 0.01259643 0.7071338 0.7320326
## Y   0.2131020 0.00011445 0.56496970 0.05150554 0.78617222 0.1100920 0.3515143
```

```
##    statusquo
## N 0.00000000
## U 0.00180032
## Y 0.00000000
```

c.

Numerical Predictor: statsquo (I'll go with "N")

"per 1-unit increase in the scale of support for the status quo, the odds of someone voting no over abstain will decrease by a factor of $e^{1.82}$, ceteris paribus." (decrease by a factor of $e^{1.82}$ ~ multiply by $e^{-1.82}$)

Dummy Variable: SexM "N"

"The odds of voting no over abstain is $e^{0.699}$ times greater for males than females, ceteris paribus."

d.

```
round(confint(fit1_chile),3)
```

```
## , , N
##
##              2.5 % 97.5 %
## (Intercept) -0.730  0.471
## age         -0.007  0.018
## sexM         0.359  1.040
## regionM     -0.711  2.394
## regionN     -0.870  0.254
## regionS     -0.166  0.835
## regionSA    -0.531  0.359
## statusquo   -2.081 -1.565
##
## , , U
##
##              2.5 % 97.5 %
## (Intercept) -0.418  0.735
## age          0.018  0.043
## sexM        -0.628  0.052
## regionM     -0.153  2.831
## regionN     -1.316 -0.158
## regionS     -0.396  0.585
## regionSA    -0.364  0.519
## statusquo    0.124  0.543
##
## , , Y
##
##              2.5 % 97.5 %
## (Intercept) -1.000  0.223
## age          0.012  0.038
## sexM        -0.458  0.250
## regionM     -0.010  3.025
## regionN     -0.502  0.663
## regionS     -0.094  0.923
```

```
## regionSA    -0.249  0.700
## statusquo    1.639  2.112
```

Numerical Predictor: statsquo "N"

"per 1-unit increase in the scale of support for the status quo, we are 95% confident that the odds of someone voting no over abstain will decrease by a factor between $e^{2.081}$ and $e^{1.565}$, ceteris paribus."

Dummy Variable: SexM "N"

"We are 95% confident that the odds of voting no over abstain is between $e^{0.359}$ and $e^{1.040}$ times greater for males than females, ceteris paribus."

Both of these seem practically significant at first look, as their effect sizes seem large enough. If further study was needed for the practicality, we could "standardize" the numerical variables.

```
summary1 <- summary(fit1_chile)
summary1
```

```
## Call:
## multinom(formula = vote ~ age + sex + region + statusquo, data = chile)
##
## Coefficients:
##   (Intercept)         age       sexM   regionM     regionN    regionS
## N  -0.1295192 0.005397365  0.6993290 0.8411402 -0.30822679 0.33434661
## U   0.1589056 0.030373441 -0.2879216 1.3387552 -0.73715898 0.09402906
## Y  -0.3884494 0.025373911 -0.1039942 1.5077273  0.08062193 0.41435538
##     regionSA  statusquo
## N -0.0860732 -1.8230660
## U  0.0771308  0.3338119
## Y  0.2254778  1.8756710
##
## Std. Errors:
##   (Intercept)         age      sexM   regionM   regionN   regionS  regionSA
## N   0.3063535 0.006410852 0.1735622 0.7920856 0.2866417 0.2554054 0.2268856
## U   0.2941819 0.006311186 0.1734192 0.7611136 0.2954569 0.2502714 0.2252501
## Y   0.3119872 0.006577450 0.1807100 0.7742880 0.2971859 0.2593320 0.2420183
##   statusquo
## N 0.1318171
## U 0.1069452
## Y 0.1207042
##
## Residual Deviance: 4245.701
## AIC: 4293.701
```

5.

$$\hat{\beta}_0' = (-0.388 - (-0.13)),\ \hat{\beta}_1' = (-0.025 - 0.006),\ \hat{\beta}_2' = (-0.104 - 0.699),\ \hat{\beta}_3' = (1.51 - 0.84)$$

$$\hat{\beta}_4' = (0.081 - (-0.31)),\ \hat{\beta}_5' = (0.41 - 0.33),\ \hat{\beta}_6' = (0.23 - (-0.09)),\ \hat{\beta}_7' = (1.88 - (-1.82))$$

```
summary1$coefficients[3,] - summary1$coefficients[1,]
```

```
## (Intercept)         age        sexM     regionM     regionN     regionS
## -0.25893016  0.01997655 -0.80332327  0.66658705  0.38884873  0.08000878
##    regionSA   statusquo
##  0.31155101  3.69873705
```

$$log(\frac{\hat{p}_{yes}}{\hat{p}_{no}}) = -0.26 + 0.02age_i - 0.803I_{sexM,i} + 0.67I_{regionM,i} + 0.39I_{regionN,i} + 0.08I_{regionS,i} + 0.31I_{regionSA,i} + 3.70stquo_i$$

$$where\ I_{sexM} \in \{0 = female, 1 = male\}\ and\ I_{regionK} \in \{0 = \text{not in region K}, 1 = \text{in region K}\},\ K = M, N, S, SA$$

Problem 2

1.

```
fit2_chile <- multinom(vote ~ age + sex + region + statusquo + age:statusquo, data = chile)
```

```
## # weights:  40 (27 variable)
## initial  value 3490.689201
## iter  10 value 2553.234153
## iter  20 value 2210.276028
## iter  30 value 2119.124632
## final  value 2118.411943
## converged
```

```
fit2_chile
```

```
## Call:
## multinom(formula = vote ~ age + sex + region + statusquo + age:statusquo,
##     data = chile)
##
## Coefficients:
##   (Intercept)        age       sexM  regionM     regionN    regionS    regionSA
## N  -0.3263155 0.01134928  0.7002179 0.864485 -0.31366574 0.33390337 -0.09228942
## U   0.0518533 0.03383230 -0.2900955 1.348366 -0.74122423 0.09227682  0.07242126
## Y  -0.1812570 0.02043307 -0.1249921 1.533334  0.08522927 0.42897067  0.22968293
##     statusquo age:statusquo
## N -2.36401743    0.01553515
## U -0.03416644    0.01109978
## Y  1.01971720    0.02362915
##
## Residual Deviance: 4236.824
## AIC: 4290.824
```

2.

There were three new parameters added, being $\beta_{j,i}, j = 1, 2, 3$. The term "age:statusquo" (or age X statusquo) represents the interaction between the two variables.

3.

```
Anova(fit2_chile)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: vote
##                LR Chisq Df Pr(>Chisq)
## age              41.55  3  4.998e-09 ***
## sex              59.56  3  7.301e-13 ***
## region           31.35 12   0.001741 **
## statusquo      1910.56  3  < 2.2e-16 ***
## age:statusquo     8.88  3   0.030972 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on our LRT statistics, we find that all of our estimates are statistically significant. The interaction between age and statusquo has the largest p-value, but still is below the threshhold of 0.05.

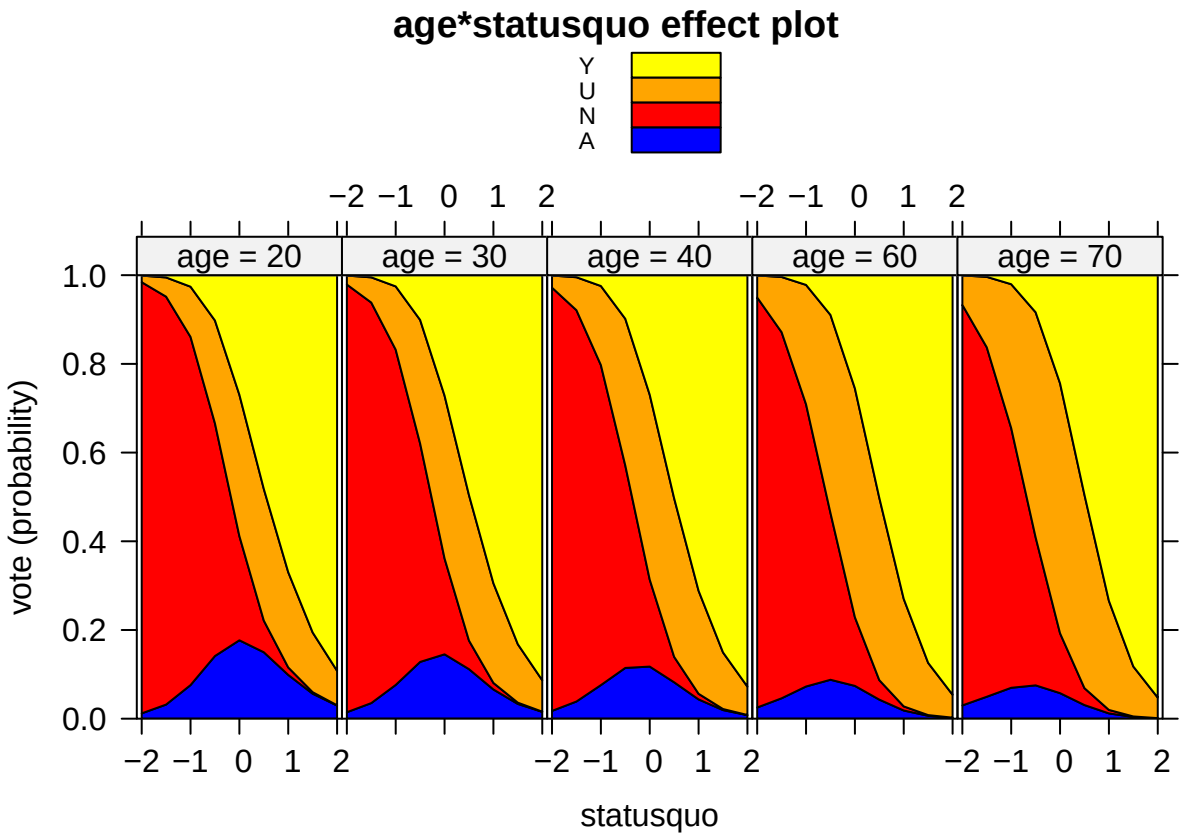
4.

"Per 1-unit increase in statusqo, the odds of voting yes over abstaining will increase, and will increase at a faster rate as age increases, ceteris paribus."

5.

```
fit2_effects <- effect("age:statusquo", fit2_chile,
                       xlevels = list(statusquo = seq(-2,2,.5)))

plot(fit2_effects, style = "stacked",
     colors = c("blue", "red", "orange", "yellow"), rug = FALSE)
```



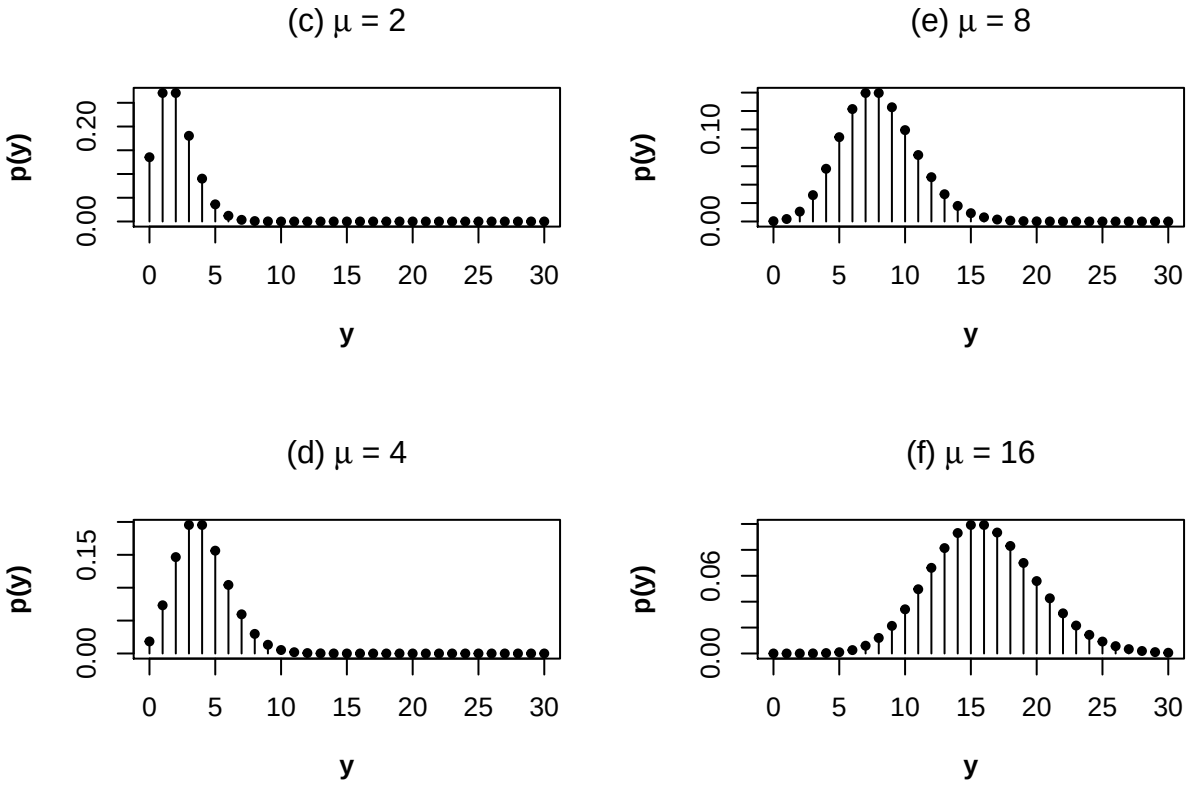
By looking at the effect displays (holding region and age constant), we can see how the effects of age and statusquo interact and change the probability of voting in a specific category. As we increase in age, we can see that as the voters get older, they are less likely to abstain. It seems that as age increases, people tend to be more active voters and those that abstain at older ages tend to be more against the statusquo. IT also appears as there is more "diversity" in active voter decisions (abstentions are not included), as voters are more "spread" between voting yes, no, and being undecided.

Problem 3

1.

```
seq_vals <- seq(0, 30, 1)
plamb_2 <- dpois(seq_vals, 2)
plamb_4 <- dpois(seq_vals, 4)
plamb_8 <- dpois(seq_vals, 8)
plamb_16 <- dpois(seq_vals, 16)
layout(matrix(c(1,2,3, 4),ncol=2))
plot(seq_vals, plamb_2, pch=20, main=expression(paste("(c) ", mu, " = 2")),
     xlab =expression(bold("y")), ylab=expression(bold("p(y)")))
segments(seq_vals, 0, seq_vals, plamb_2)
plot(seq_vals, plamb_4, pch=20, main=expression(paste("(d) ", mu, " = 4")),
     xlab =expression(bold("y")), ylab=expression(bold("p(y)")))
segments(seq_vals, 0, seq_vals, plamb_4)
```

```
plot(seq_vals, plamb_8, pch=20, main=expression(paste("(e) ", mu, " = 8")),
     xlab =expression(bold("y")), ylab=expression(bold("p(y)")))
segments(seq_vals, 0, seq_vals, plamb_8)
plot(seq_vals, plamb_16, pch=20, main=expression(paste("(f) ", mu, " = 16")),
     xlab =expression(bold("y")), ylab=expression(bold("p(y)")))
segments(seq_vals, 0, seq_vals, plamb_16)
```



Note: I was able to make the graps have 1x1 aspect ratios like in the slides, but the graphs were too small in the pdf output for some reason. If you would like to see the code so they are exactly the same (besides the size), I will be happy to provide it.

2.

```
# mu = 2
ppois(10, 2, lower.tail = TRUE) - ppois(5,2, lower.tail = TRUE)
```

```
## [1] 0.0165553
```

```
# mu = 4
ppois(10, 4, lower.tail = TRUE) - ppois(5,4, lower.tail = TRUE)
```

```
## [1] 0.2120298
```

```r
# mu = 8
ppois(10, 8, lower.tail = TRUE) - ppois(5,8, lower.tail = TRUE)
```

```
## [1] 0.6246497
```

```r
# mu = 16
ppois(10, 16, lower.tail = TRUE) - ppois(5,16, lower.tail = TRUE)
```

```
## [1] 0.07601223
```

$P(X \in [5, 10])$, where $X \sim Pois(2)$, is 0.0166

$P(X \in [5, 10])$, where $X \sim Pois(4)$, is 0.212

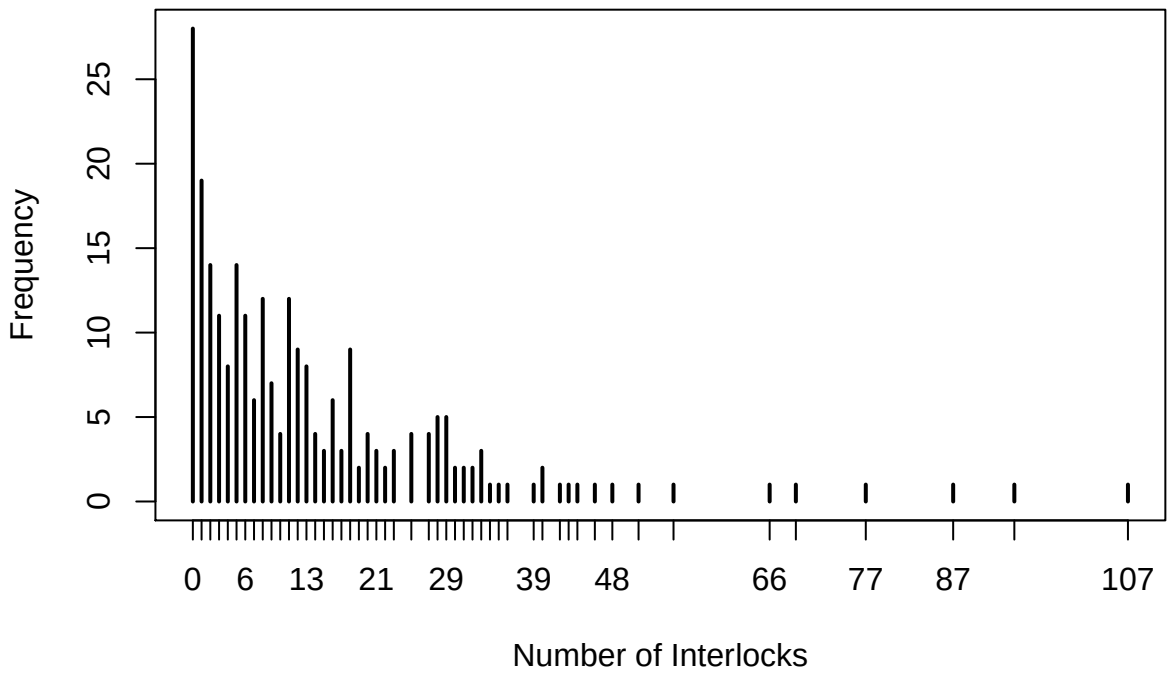
$P(X \in [5, 10])$, where $X \sim Pois(8)$, is 0.625

$P(X \in [5, 10])$, where $X \sim Pois(16)$, is 0.076

3.

We are working with count data for the number of interlocks. Clearly right-skewed (wow. this looks familiar at the surface level).

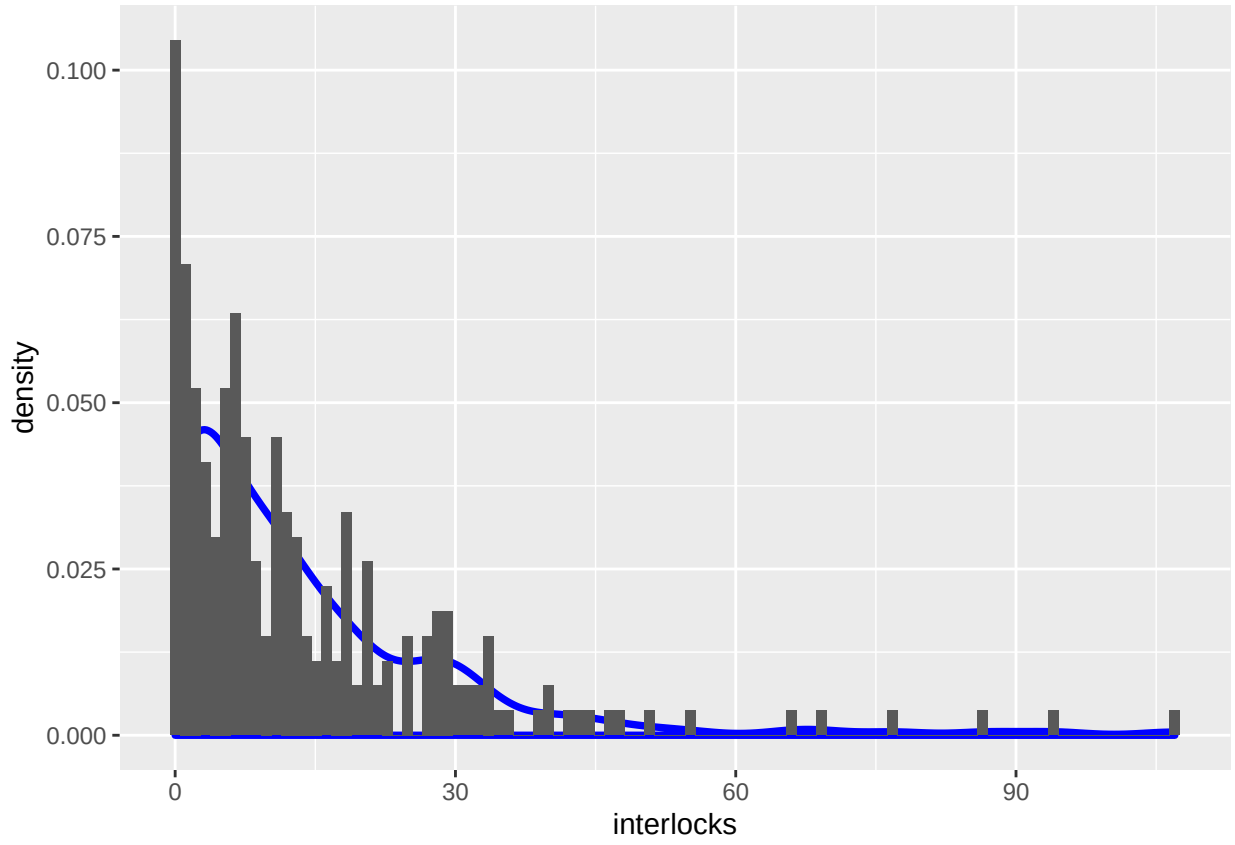
```r
plot(table(Ornstein$interlocks), xlab="Number of Interlocks", ylab="Frequency")
```



Since we want to find $\hat{\mu}$ using MLE, we need to find the value of μ that maximizes the likelihood, given our distribution. We can get an idea by looking at the the density plot.

```
ggplot(data = ornstein, aes(x=interlocks, y = ..density..))+
  geom_density(bins = 100, col = "blue", size = 1.3) +
  geom_histogram(bins = 100)
```

```
## Warning: Ignoring unknown parameters: bins
```



Now, we could estimate μ by simply finding the mean of the distribution as it is the most likely estimate given our data. That value you would be (rounding to nearest whole number): 14

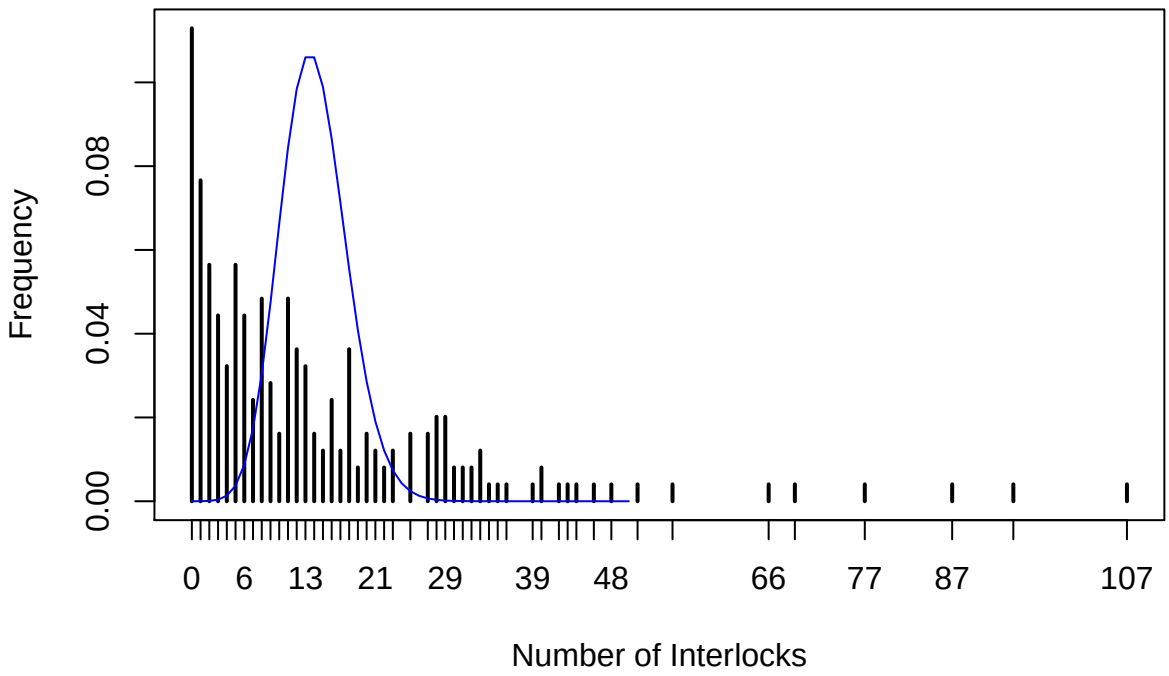
```
mean(ornstein$interlocks)
```

```
## [1] 13.58065
```

We can then plot the poisson distribution given $\hat{\mu}$ (notice the scale for y-axis).

```
seq_vals <- seq(0, 50)
estimates <- dpois(seq_vals, 14)
df <- data.frame(seq_vals, estimates)
colnames(df) <- c("values", "estimates")

plot(table(Ornstein$interlocks)/length(Ornstein$interlocks), xlab="Number of Interlocks", ylab="Frequen
lines(df$values, df$estimates, col ="blue")
```



This doesn't seem like a great model, as there is a lot more data closer to 0 and the distribution is right-skewed, meaning the mean of our data is highly affected by the skew.