StatsHw8

Trevor Freeland

May 10, 2018

```
contraception <- read.csv("http://math.carleton.edu/Chihara/Stats345/Contraception.csv")
campuscrime <- read.csv("http://math.carleton.edu/Chihara/Stats345/CampusCrime.csv")
suicides <- read.csv("http://math.carleton.edu/Chihara/Stats345/Suicides.csv")
ticks <- read.csv("http://math.carleton.edu/Chihara/Stats345/GrouseTicks.csv")
```

```
library(tidyverse)
library(lme4)
library(stargazer)
library(plyr)
library(pander)
```

64

(A)

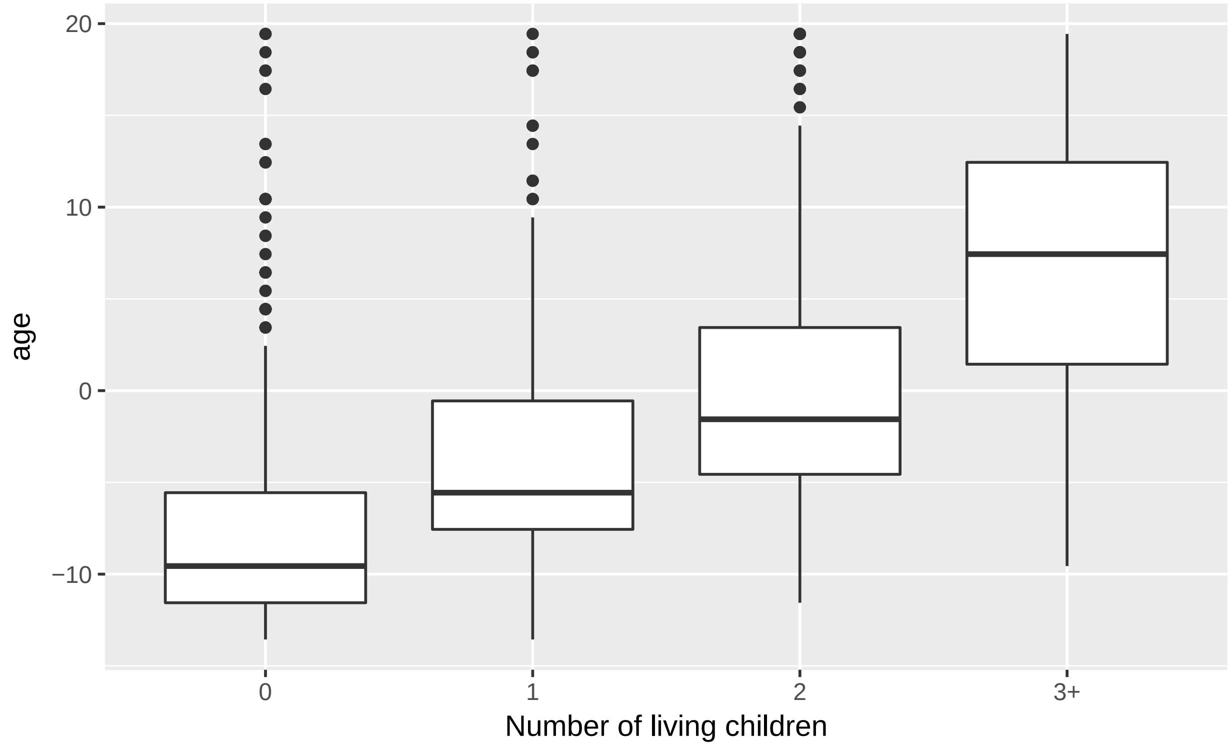
From our summary table we can see that there appears to be a lot more woman in rural regions in this study than in metropolitan areas. We can also see that about 60% (1175/(1175+759)) of the women were not using contreception at the time of this survey. In our first boxplot below we can see that there is correlation between age and number of living children. This intuitively makes sense, because younger women have had less time to have children compared to older women. Then in our second boxplot we can see that on average there doesn't appear to be a large age difference in contraceptive use, however there does appear to be more variablitly in ages of those that did not use contraceptions than those that did.

```
pander(summary(contraception))
```

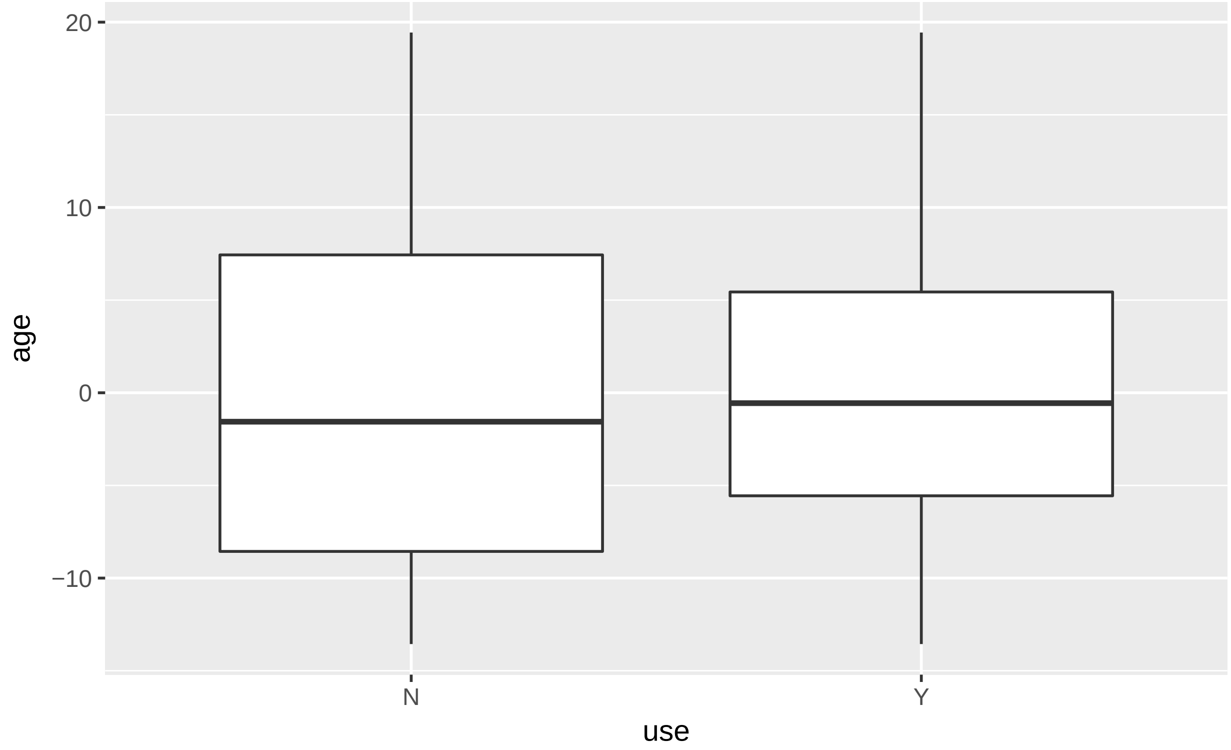
Table 1: Table continues below

woman	district	use	livch	age	region
Min. : 1.0	Min. : 1.00	N:1175	0 :530	Min. :-13.560000	Metropolitan: 562
1st Qu.: 484.2	1st Qu.:14.00	Y: 759	1 :356	1st Qu.: -7.559900	Rural :1372
Median : 967.5	Median :29.00	NA	2 :305	Median : -1.559900	NA
Mean : 967.5	Mean :29.35	NA	3+:743	Mean : 0.002198	NA
3rd Qu.:1450.8	3rd Qu.:45.00	NA	NA	3rd Qu.: 6.440000	NA
Max. :1934.0	Max. :61.00	NA	NA	Max. : 19.440000	NA

```
ggplot(data = contraception, aes(x = livch, y = age)) + geom_boxplot() + xlab("Number of living childre
```



```
ggplot(data = contraception, aes(x = use, y = age)) + geom_boxplot()
```



(B)

Level 1: LogOdds(Use) $= a_i + b(age) + c(livch1) + d(livch2) + e(livch3+)$

Level 2: $a_i = \alpha_0 + \alpha_1(Rural) + \mu_i$ where μ_i~N(0, σ_1^2)

Estimated LogOdds(Use) = -.96 + 1.11(livch1) + 1.38(livch2) + 1.35(livch3+) -.03(age) -.73(Rural)

(C)

The odds of a woman using contraception who has one living child is 3.03 (`exp(1.11)`) times greater than the odds of a woman using contraception that has zero living children, holding everything else constant.

The odds of a woman using contraception who has two living child is 3.97 (`exp(1.38)`) times greater than the odds of a woman using contraception that has zero living children, holding everything else constant.

The odds of a woman using contraception who lives in a rural region is .48 (`exp(-.73)`) times smaller than the odds of a woman using contraception who lives in a metropolitan region, holding everything else constant.

```
contraception.lmer1 <- glmer(use~(livch+age+region)^2 + (1|district), data = contraception, family = bi
summary(contraception.lmer1)

#Do we need interactions?
contraception.lmer3 <- glmer(use~livch + age + region + (1|district), data = contraception, family = bi
anova(contraception.lmer3, contraception.lmer1) #Not really.
summary(contraception.lmer3)
```

65

(A)

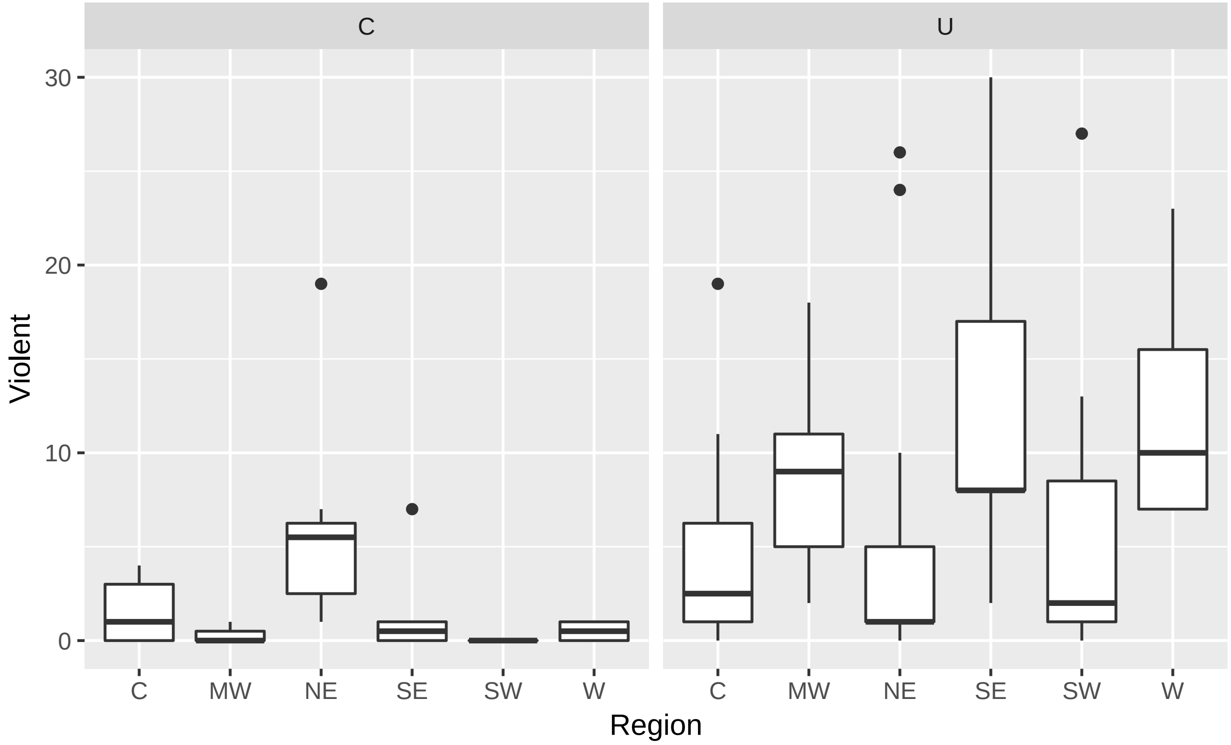
In our table of summary statistics we can see that there are possibly some outliers in regards to property crimes and total crimes. The maxes are well above not only the mean and median, but also well above the 3rd quartile. Also scaling might an issue for pretty much all of our numeric variables since the differences between the maxes and mins are very large.

In our plot below we can see that the different regions all seems to have different distributions with the number of violent crimes. Also there appears to be a significant different between Colleges and Universities, and the effects seem to vary based on the different regions, which implies that there might be some interactions that we need to be accounting for.

```
pander(summary(campuscrime))
```

Region	Type	Enrollment	Property	Violent	Total
C :17	C:28	Min. : 540	Min. : 8.0	Min. : 0.000	Min. : 9.0
MW:10	U:53	1st Qu.: 4638	1st Qu.: 64.0	1st Qu.: 1.000	1st Qu.: 65.0
NE:21	NA	Median :11321	Median : 120.0	Median : 3.000	Median : 127.0
SE:15	NA	Mean :13899	Mean : 240.8	Mean : 5.938	Mean : 246.7
SW:10	NA	3rd Qu.:22396	3rd Qu.: 293.0	3rd Qu.: 8.000	3rd Qu.: 301.0
W : 8	NA	Max. :46597	Max. :1293.0	Max. :30.000	Max. :1311.0

```
ggplot(data = campuscrime, aes(x = Region, y = Violent)) + geom_boxplot() + facet_wrap(~Type)
```



(B)

```
model1 <- glm(Violent~Region+Type, data = campuscrime, family = poisson)
```

(C)

With a dispersion parameter of 7.89 there is definitely something going on with our model.

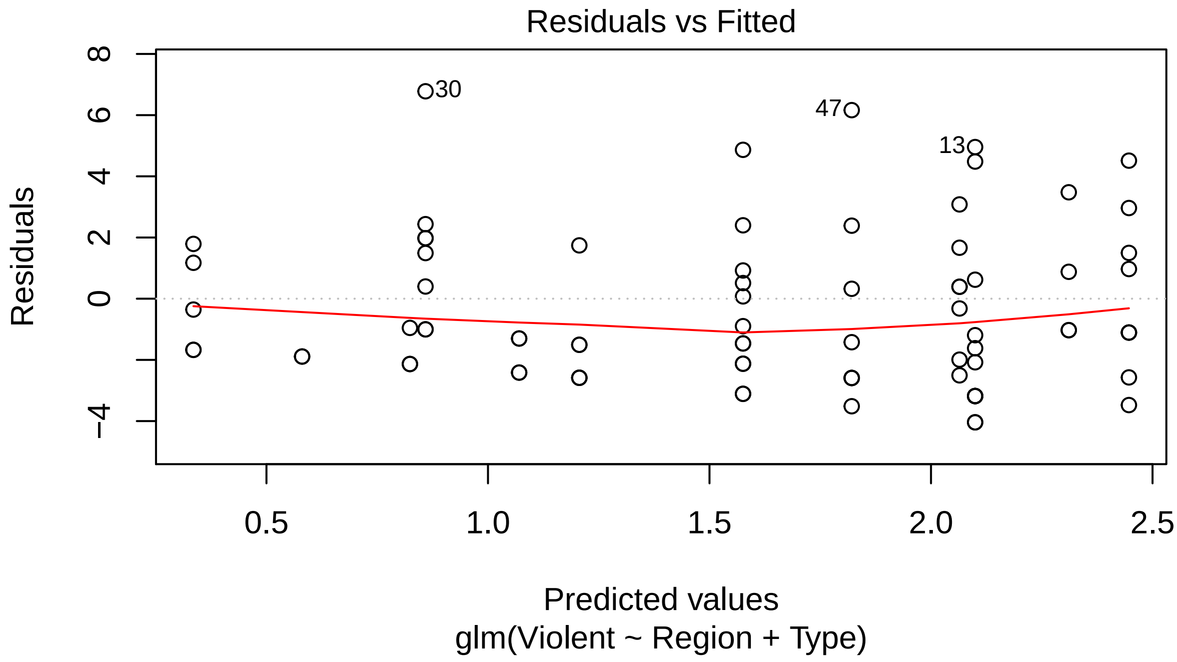
```
sum(residuals(model1, type = "pearson")^2)/model1$df.residual
```

```
[1] 7.890309
```

(D)

Even after removing observations that had pearson residuals greater than 4 or less than -4 and re-running our model we still get a dispersion parameter of greater than 4, which is smaller than before, but there is still something going on with our model.

```
plot(model1, which=1) #Definitely outliers.
```



```
index <- which(abs(resid(model1, type = "pearson")) > 4) #grabbing any residuals greater than 4
campuscrimenew <- campuscrime[-index,]
model2 <- update(model1, data = campuscrimenew)
sum(residuals(model2, type = "pearson")^2)/model2$df.residual #Smaller, but still overdispersion
```

66

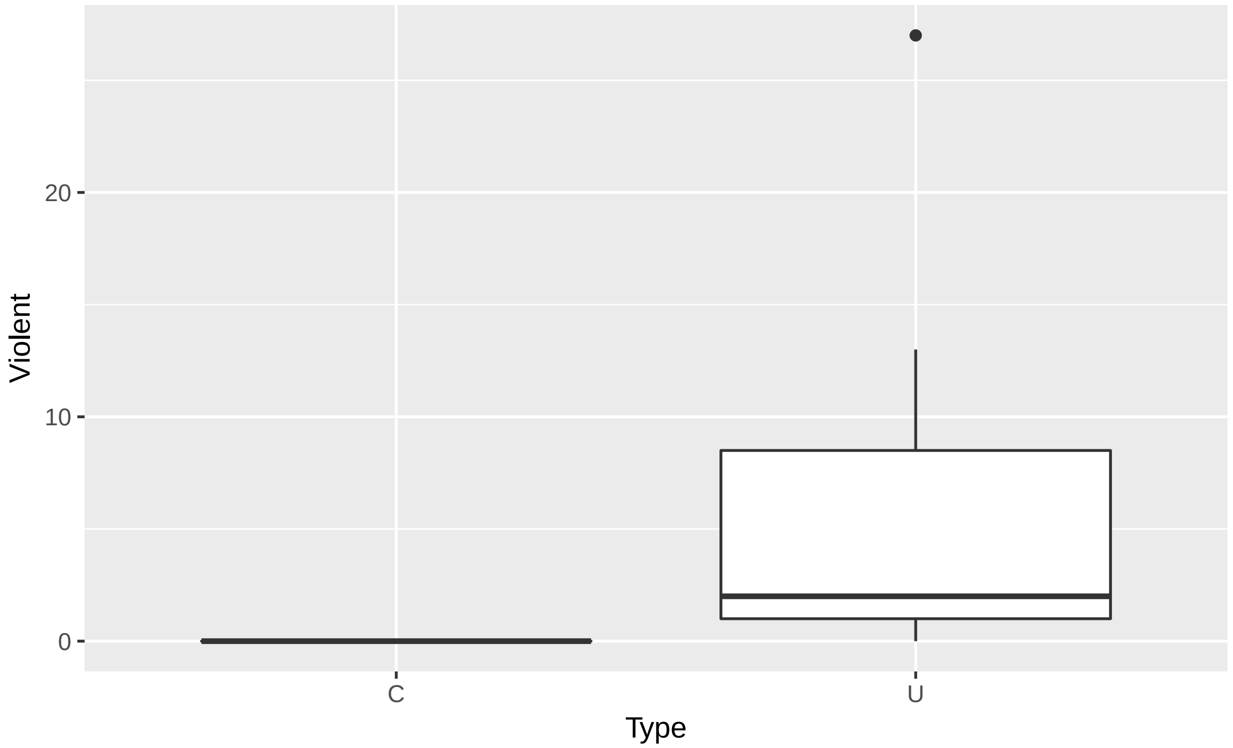
(A)

There appears to be something werid going on in the southwest, we get a standard deviation of 902 for both the region and the interaction term. Some of the other interactions appear significant but not all of them.

(B)

In the graph right below of violent crimes by type of school in the SW we can see that there is a lot of variability in the university type, but any college in the SW has 0 violent crimes, which is causing issues in our model most likely, because that is giving us a definitive answer, if the school is a college in the SW, there will definitively be 0 violent crimes(Based on our data).

```r
ggplot(data = filter(campuscrime, Region == "SW")) + geom_boxplot(aes(x = Type, y = Violent))
```



(C)

After combining the West and SouthWest into one Region we have gotten rid of that extremely large standard deviation term that we had from before.

(D)

Our dispersion parameter is now 3.34, which is again getting smaller, but it is still significantly larger than 1 so we need to account for it.

```r
model3 <- glm(Violent~Region*Type, data = campuscrimenew, family = poisson)
summary(model3)
```

```r
campuscrimenew$Region <- mapvalues(campuscrimenew$Region, c("SW", "W"), c("Region2", "Region2"))
model4 <- update(model3, data = campuscrimenew)
summary(model4)
```

```r
sum(residuals(model4, type = "pearson")^2)/model4$df.residual #Smaller still, but still overdispersion
```

67

(A)

```r
model5 <- update(model4, family = quasipoisson)
summary(model5)
```

(B)

After taking out the interaction terms, it suggests that only Region of SE is significant along with the Type of university.

```r
model6 <- glm(Violent~Region+Type, data = campuscrimenew, family = quasipoisson)
summary(model6)
```

(C)

The anova test suggests that we need the interactions terms in our model.

```r
anova(model5, model6, test = "Chisq")
```

(D)

When looking at 95% confidence intervals for all of the coefficients, we have lower bounds in the negatives and upper bounds above 0, which means that 0 is included in all of the confidence intervals, so it is possible that some of the effects could be 0.

```r
confint(model5)
```

(E)

```r
campuscrimenew$Region3 <- as.factor(ifelse(campuscrimenew$Region == "NE", "NE", "Other"))
```

(F)

Final Model: Log(μ) = 1.42 - 1.42(RegionOther) - .52(University) + 2.41(RegionOther:University)

The average number of violent crimes at a college in the south east is `exp(1.42)` = 4.14.

The average number of violent crimes is .24 (`exp(-1.42)`) times smaller than for campuses in the South East, holding everything else constant.

The average number of violent crimes is .59 (`exp(-.52)`) times smaller for universities in the South East compared to colleges in the South East, holding everything else constant.

The average number of violent crimes is 11.1 (`exp(2.41)`) times greater for a university in a region other than the south east than for a college in a region other than the south east, holding everything else constant.

```r
model7 <- glm(Violent~Region3*Type, data = campuscrimenew, family = quasipoisson)
summary(model7)
```

68

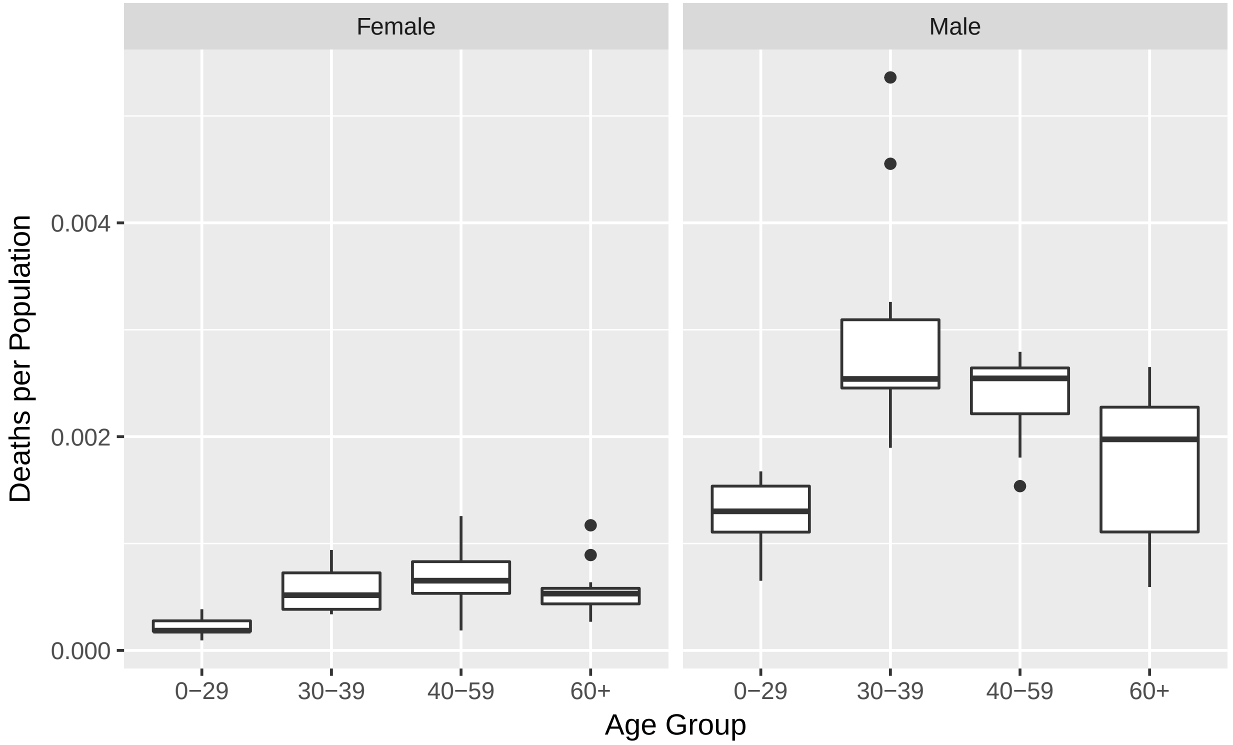
(A)

In our table below we can see some summary stats for our numeric variables. There appears to be a very large range for our Population variable, especially compared to the deaths variable. In our plot below we can see that there is a difference in suicides between the age groups, and also that there appears to be a relationship between sex and suicides. This plot shows that there also might be an interaction between sex and age group, because the effects don't seem to simply shift up going from Female to Male, depending on the age group the effect seems to be slightly different.

```
suicides$sex <- as.factor(ifelse(suicides$sex == 0, "Female", "Male"))
suicides <- suicides %>% select(-c(smr,expected))
table <- rbind(summary(suicides$pop), summary(suicides$death), summary(suicides$death/suicides$pop))
row.names(table) <- c("Pop", "Deaths", "Deaths/Pop")
pander(table)
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Pop	2434	13192	23878	34512	46879	210422
Deaths	1	9	21	36.77	52.75	198
Deaths/Pop	9.409e-05	0.0005042	0.000947	0.001292	0.002121	0.00536

```
ggplot(data = suicides, aes(x = age, y = death/pop)) + geom_boxplot() + facet_wrap(~sex) + xlab("Age Gr
```



(B)

The estimate for the dispersion parameter is 1.85, which indicates we might not be accounting for something in our model.

```
suicide.glm <- glm(death~Region + sex + age + offset(log(pop)), data = suicides, family = poisson)
summary(suicide.glm)
sum(residuals(suicide.glm, type = "pearson")^2)/suicide.glm$df.residual
```

(C)

Our dispersion parameter lowered from 1.85 to 1.41 after adding the interactions between sex and age.

Our estimated Model: log(mean deaths) = -8.14 -.39(Dublin) -.69(EHB-Dub) -.43(Galway) -.15(Lim.) -.39(Mid HB) -.29(MWHB - Lim.) -.35(NEHB) -.36(NWHB) -.15(SEHB - Wat.) -.24(SHB-Cork) -.55(Waterf) -.29(WHB-Gal.) + 1.80(Male) + .98(age30-39) + 1.24(age40-59) + .95(age60+) -.24(Male30-39) -.63(Male40-59) -.59(Male60+)

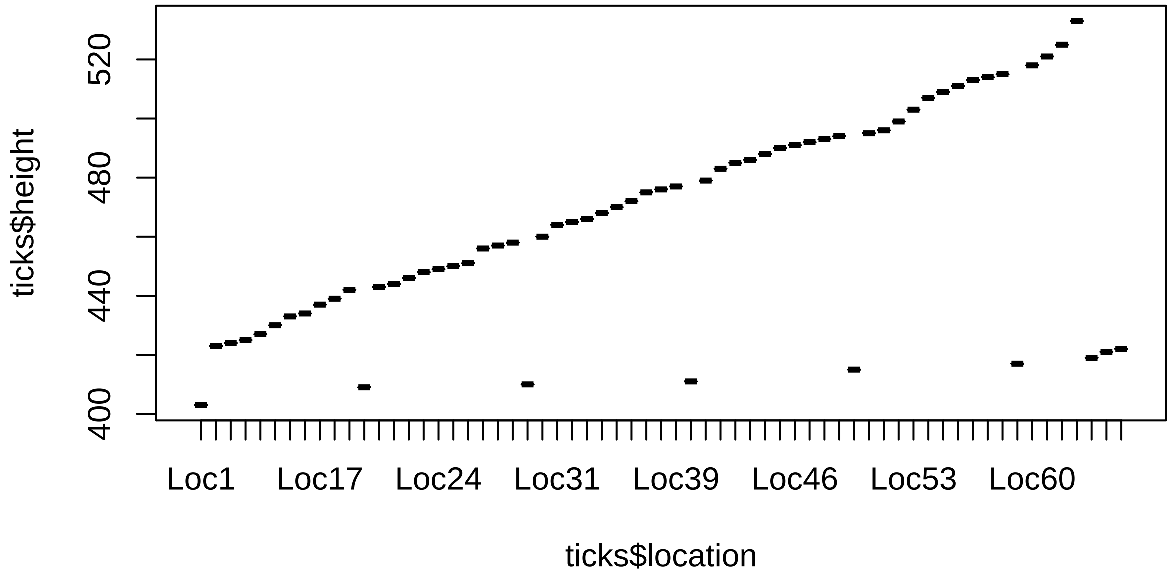
```
suicide.glm2 <- glm(death~Region + sex * age + offset(log(pop)), data = suicides, family = poisson)
summary(suicide.glm2)
sum(residuals(suicide.glm2, type = "pearson")^2)/suicide.glm2$df.residual
```

69

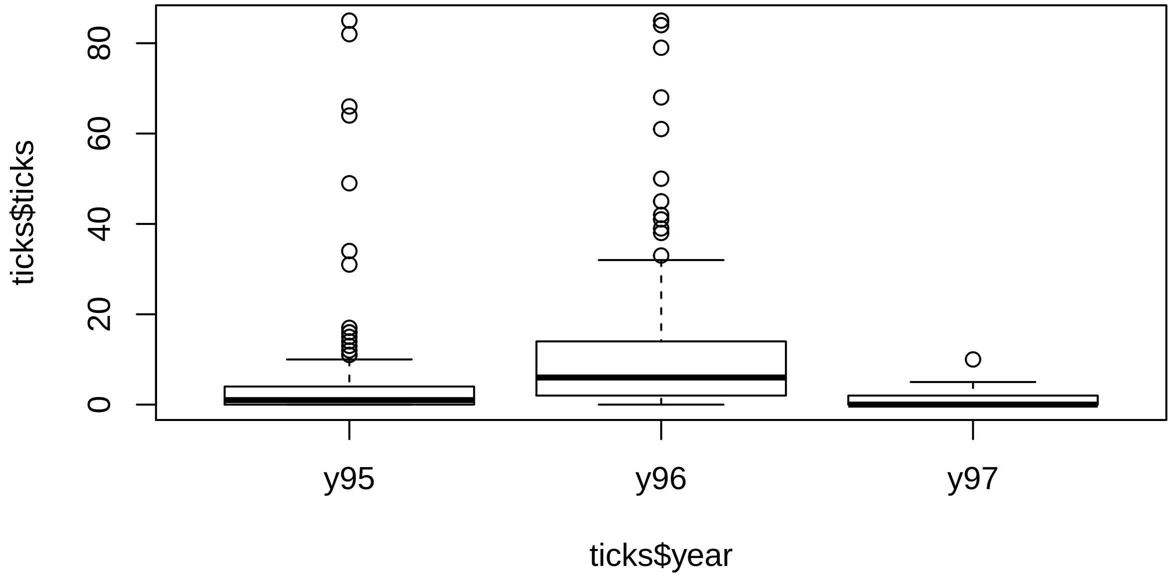
(A)

In the first plot below we plot the height against the location. From this plot we can see that each height value pairs up with a single location value and vice versa. There are also 63 unique height values and 63 unique location values so it appears that those two are telling us the same information. When looking at the number of ticks and the year in which the chick was born we can see that there is a difference in the year that the chick was born, specifically that y96 seems to have significantly more ticks than the other years.

```
ticks$ht <- as.factor(ticks$height)
plot(ticks$height~ticks$location)
```



```
plot(ticks$ticks~ticks$year)
```



(B)

Our model is the number of ticks against the year and height, with a random intercept for the brood in the location, and at the location level.

log(mean#Ticks) = 11.35 + 1.17(y96) -.98(y97) - 2.35(Height)

```
ticks$ht <- ticks$height/100
ticks.glmer <- glmer(ticks~year + ht + (1|location/brood), data = ticks, family = poisson)
summary(ticks.glmer)
fixef(ticks.glmer)
```

(C)

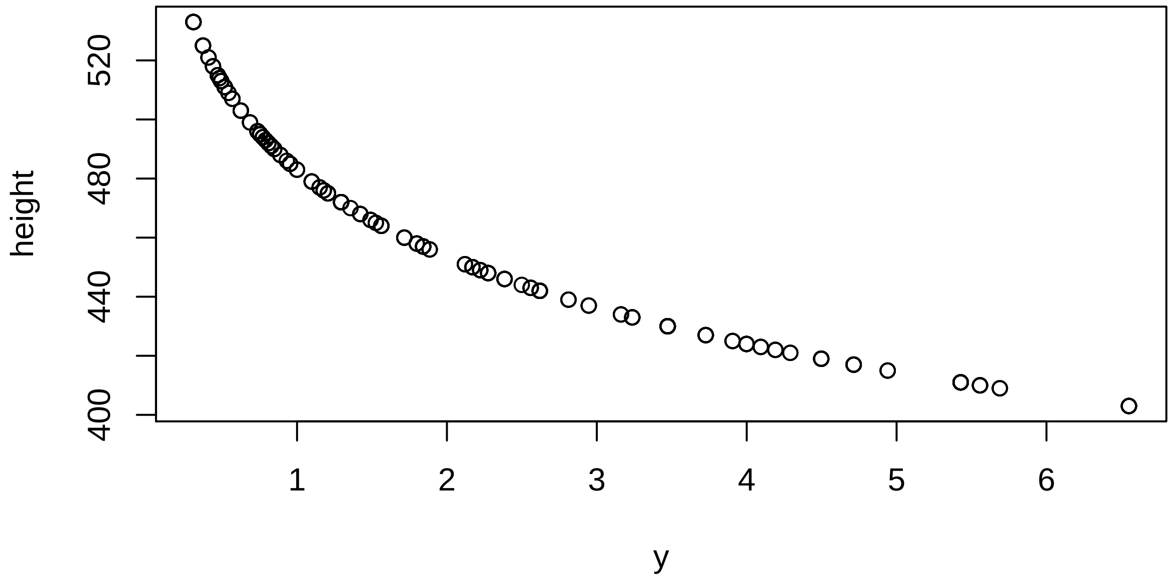
64% is due to the brood while 36% is due to the location.

```
loc <- .3296/(.3296+.5924)
brood <- .5924/(.3296+.5924)
```

(D)

I don't know how to get the estimated mean number of ticks against height. It also still seems odd to have height in our model since it is perfectly correlated to the location. Also we haven't done hierarchical poisson models so my whole model may be incorrect.

```
ticks$y <- exp((11.35 -2.35*ticks$ht))
plot(height~y, data = ticks)
```


height
y