Draper HW 7

Matthew Draper

May 17, 2019

```
library(foreign); library(MASS); library(pscl); library(stargazer)
library(vcd); library(AER); library(countreg); library(ggplot2); library(sjPlot)
```

For this exercise you will need to load the Muller & Seligson file, Msrepl87.asc.

```
ms<-read.table("Msrepl87.asc", header=TRUE)
ms$sanctions <- (ms$sanctions70 + ms$sanctions75)/2
ms<-ms[,c(1,3,18,19,23,24)]
ms<-as.data.frame(na.omit(ms))
    #colClasses=c("character",rep("numeric",22)))
    #rownames(ms) <- ms$country
```

1. Using the raw data, construct a model that uses the log of deaths75 as a dependent variable. Select up to four independent variables of your choice. Since the DV data contain zeros, you will have to add a constant before taking the logarithm. Try at least 4 different values for c at different orders of magnitude and compare your statistical results in a table and graphically. What did you learn?

```
c <- .01
ms$logdeaths75 <- log(ms$deaths75 + c)
m1 <- glm(logdeaths75 ~ sanctions + giniland + landless + energypc, data = ms)
m1c<-coefficients(m1)

c <- .1
ms$logdeaths75 <- log(ms$deaths75 + c)
m2 <- glm(logdeaths75 ~ sanctions + giniland + landless + energypc, data = ms)
m2c<-coefficients(m2)

c <- 1
ms$logdeaths75 <- log(ms$deaths75 + c)
m3 <- glm(logdeaths75 ~ sanctions + giniland + landless + energypc, data = ms)
m3c<-coefficients(m3)

c <- 10
ms$logdeaths75 <- log(ms$deaths75 + c)
m4 <- glm(logdeaths75 ~ sanctions + giniland + landless + energypc, data = ms)
m4c<-coefficients(m4)
```

Adding a constant moves the mean, and more seriously, it artificially reduces the variance of the data. The larger the value of c, the lower the variance will be. Adding arbitrary constants can alter model fit and interpretation. Indicators of model fit shift dramatically simply with the addition of a constant. In the residuals vs. fitteed and Q-Q plots below, we observe clear evidence of non-normality and heteroskedasticity.

```
q1<-rbind(m1c,m2c,m3c,m4c)
```

```
stargazer(q1)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, May 17, 2019 - 5:28:09 PM

Table 1:

	(Intercept)	sanctions	giniland	landless	energypc
m1c	-0.200	0.019	1.621	0.064	-0.0002
m2c	0.698	0.016	1.211	0.056	-0.0002
m3c	1.686	0.013	0.758	0.048	-0.0002
m4c	2.999	0.009	0.228	0.039	-0.0001

```
stargazer(m1,m2,m3,m4)
```

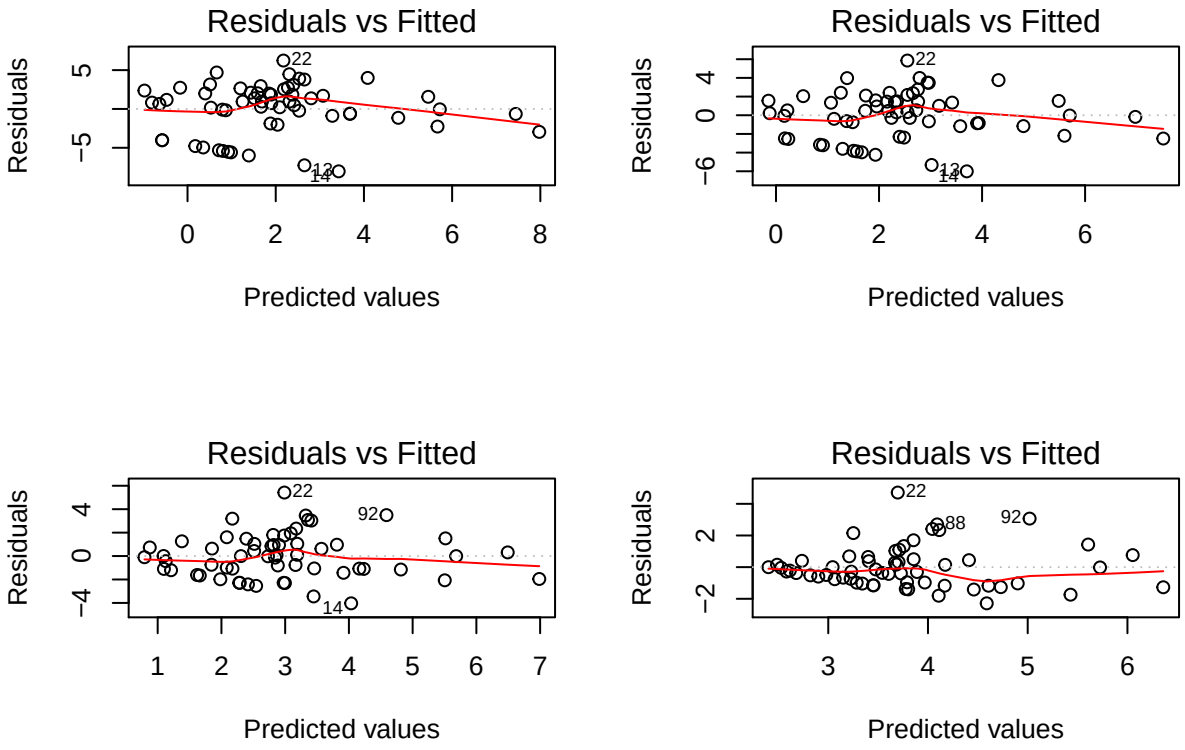
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, May 17, 2019 - 5:28:09 PM

Table 2:

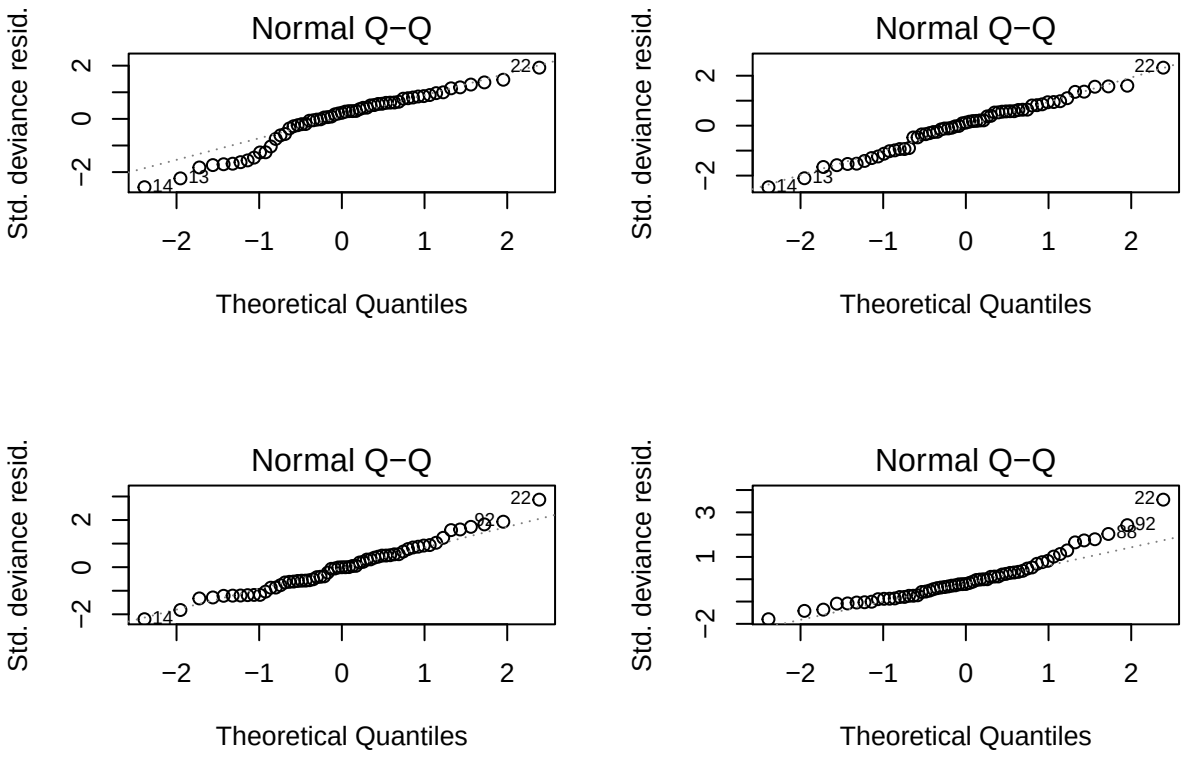
	Dependent variable:			
	logdeaths75			
	(1)	(2)	(3)	(4)
sanctions	0.019***	0.016***	0.013***	0.009***
	(0.006)	(0.004)	(0.003)	(0.002)
giniland	1.621	1.211	0.758	0.228
	(2.460)	(1.916)	(1.437)	(1.003)
landless	0.064	0.056	0.048	0.039*
	(0.051)	(0.039)	(0.030)	(0.021)
energypc	−0.0002	−0.0002	−0.0002	−0.0001
	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Constant	−0.200	0.698	1.686*	2.999***
	(1.646)	(1.282)	(0.962)	(0.671)
Observations	59	59	59	59
Log Likelihood	−153.354	−138.603	−121.648	−100.444
Akaike Inf. Crit.	316.709	287.207	253.296	210.888
Note:				*p<0.1; **p<0.05; ***p<0.01

We observe that Model 1 has a negative intercept, and all four linear models generate non-integer data. These diagnostics should lead us to consider an alternative model of the data-generating process.

```
par(mfrow=c(2,2))
plot(m1,1)
plot(m2,1)
plot(m3,1)
plot(m4,1)
```



```
par(mfrow=c(2,2))
plot(m1,2)
plot(m2,2)
plot(m3,2)
plot(m4,2)
```

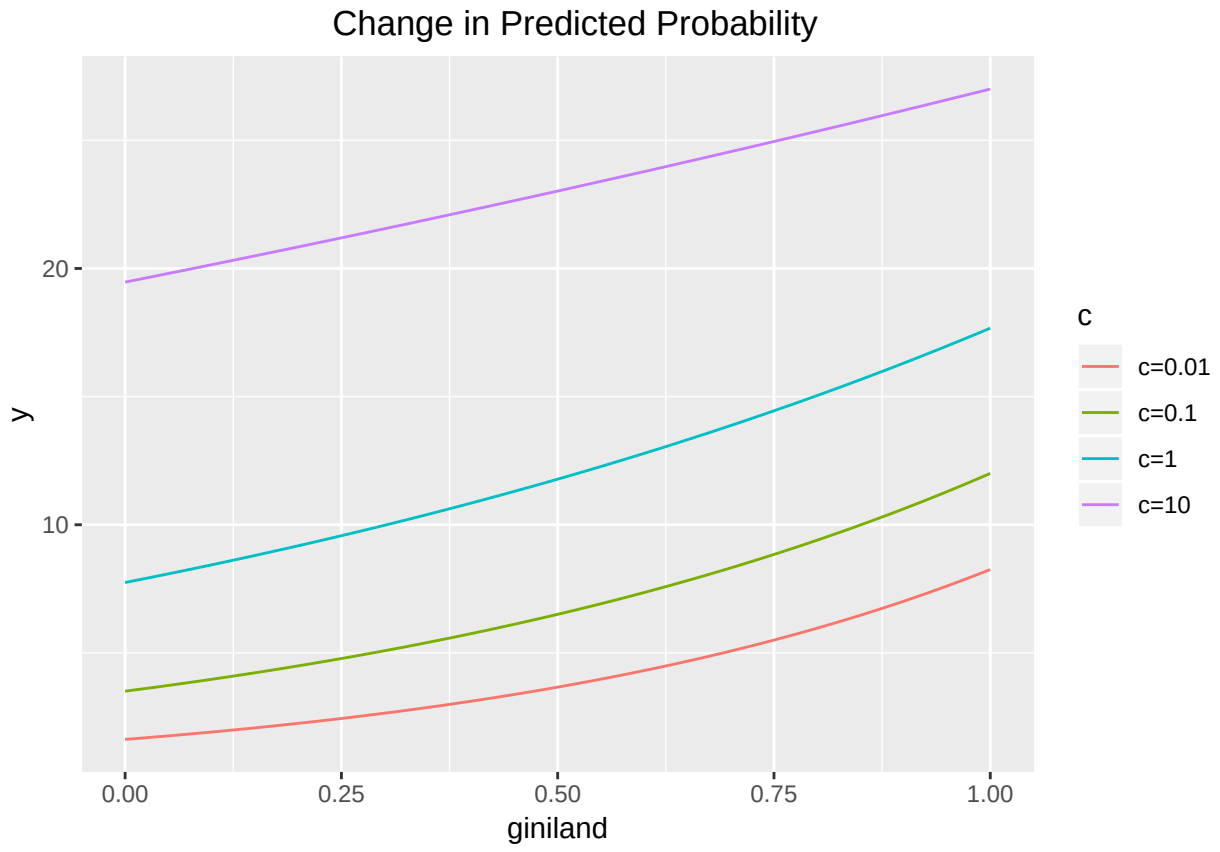


In the residuals vs. fitted plots, we observe a wedge/fan shape in the data, indicating that the homoskedasticity assumption has been violated. In addition, the Q-Q plots indicate that the data are not normally distributed. Finally, Model 1 has a negative intercept, and all four linear models generate non-integer data. These diagnostics should lead us to consider an alternative model of the data-generating process.

I will now simulate new data and plot the change in predicted probability. The scenario I would like to explore is one where the 'giniland' variable is allowed to vary across its full range while the three other variables in the model are held at their central tendencies.

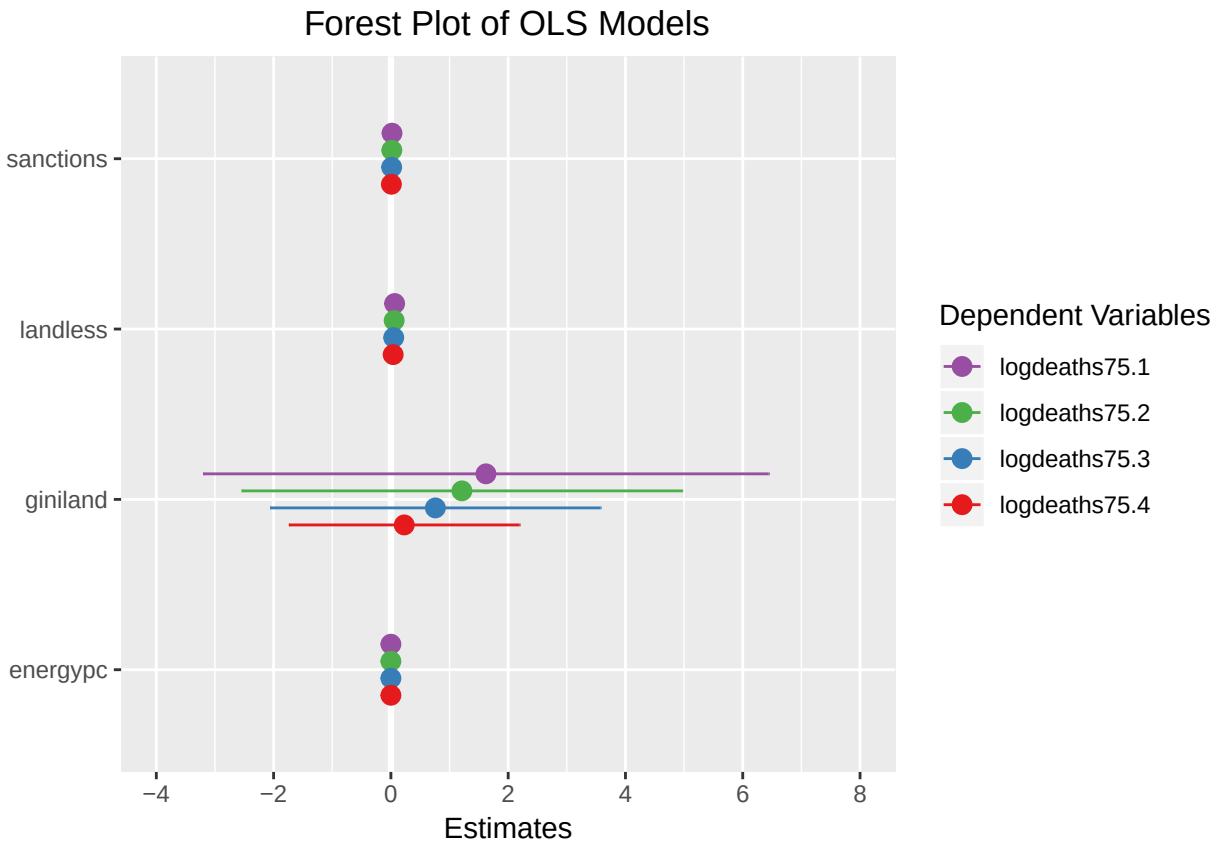
```
x <- data.frame(sanctions=rep(median(ms$sanctions),101),
                giniland=seq(0,1,by=0.01),
                landless=rep(median(ms$landless,na.rm = TRUE),101),
                energypc=rep(median(ms$energypc,na.rm = TRUE),101))
pred<-predict(m1,x)
d1=exp(pred)-.01
pred<-predict(m2,x)
d2=exp(pred)-.1
pred<-predict(m3,x)
d3=exp(pred)-1
pred<-predict(m4,x)
d4=exp(pred)-10
data<-data.frame(x=rep(x$giniland,4),y=c(d1,d2,d3,d4),c=c(rep('c=0.01',101),rep('c=0.1',101),rep('c=1',

plot1<-ggplot(data,aes(x=x,y=y,colour=c))+geom_line(aes(group=c))
plot1 + xlab("giniland") + labs(title = '                          Change in Predicted Probability')
```



As we increase the value of c, we observe both a shift in the y-intercept and a flattening of the curve. By adding a constant to each observation, we have shifted the mean by the value of our constant.

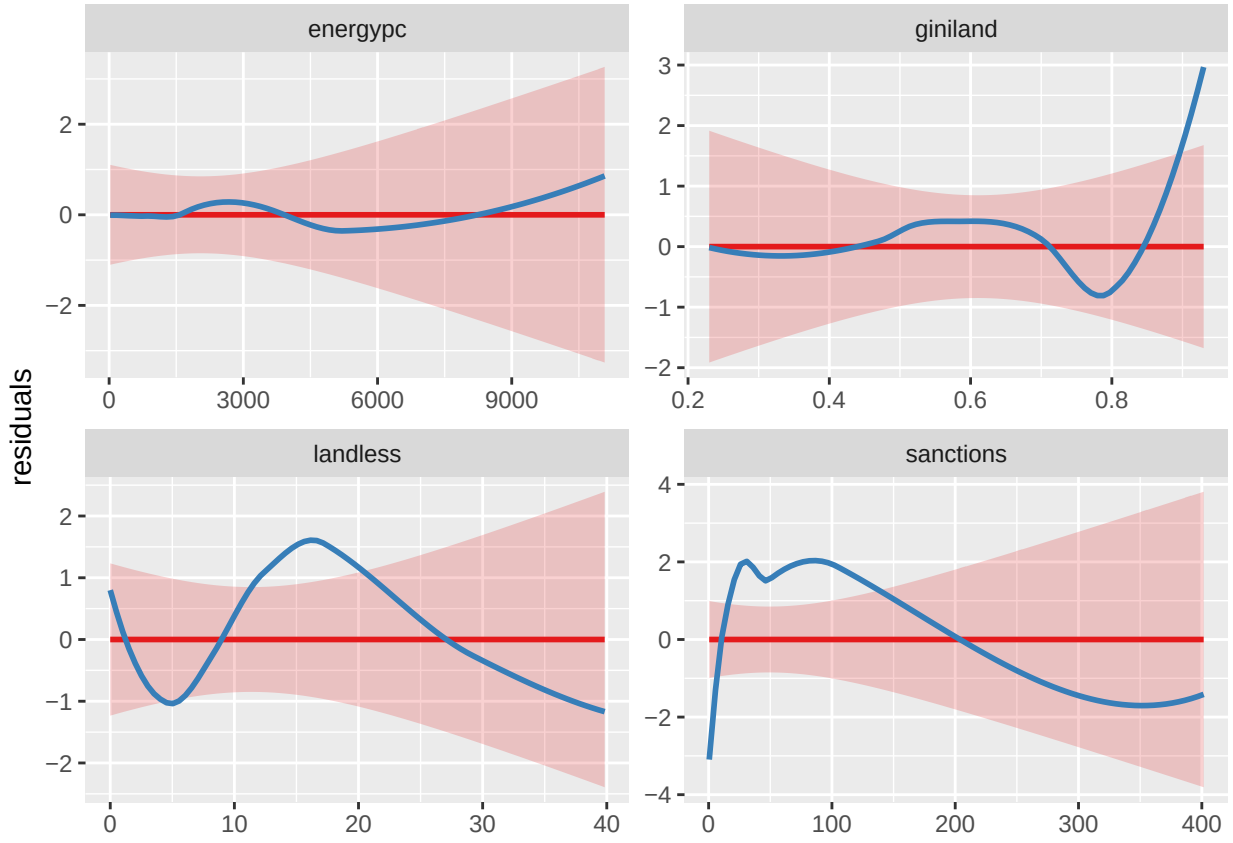
```
plot_models(m1,m2,m3,m4, title = '                    Forest Plot of OLS Models')
```



We observe significance for all variables except 'giniland'.

Here I give limited model diagnostics for m1. Remaining models are omitted to save space, but may be enabled by un-commenting the code below.

```
#plot_model(m1, type = "pred")
#plot_model(m1, type = "diag")
plot_model(m1, type = "resid")
```




```
#par(mfrow=c(2,2))
#plot_model(m1, type = "pred")
#plot_model(m2, type = "pred")
#plot_model(m3, type = "pred")
#plot_model(m4, type = "pred")

#plot_model(m1, type = "resid")
#plot_model(m2, type = "resid")
#plot_model(m3, type = "resid")
#plot_model(m4, type = "resid")

## t <- data.frame(m1$residuals,ms$logdeaths75)
## plot(m1$residuals,ms$logdeaths75)
## m1$residuals
## Graph residuals (x) vs. log deaths (y)
```

Information loss is clearly visible in the flattening of the residual curves.

2. Using the same specification of the linear predictor as in part (1) but omitting any log transformations of the dependent variable, specify and estimate a Poisson model. Is there any evidence of overdispersion?

```
p1 <- glm(deaths75 ~ sanctions + giniland + landless + energypc, data = ms, family = 'poisson')
```

```
stargazer(p1)
```

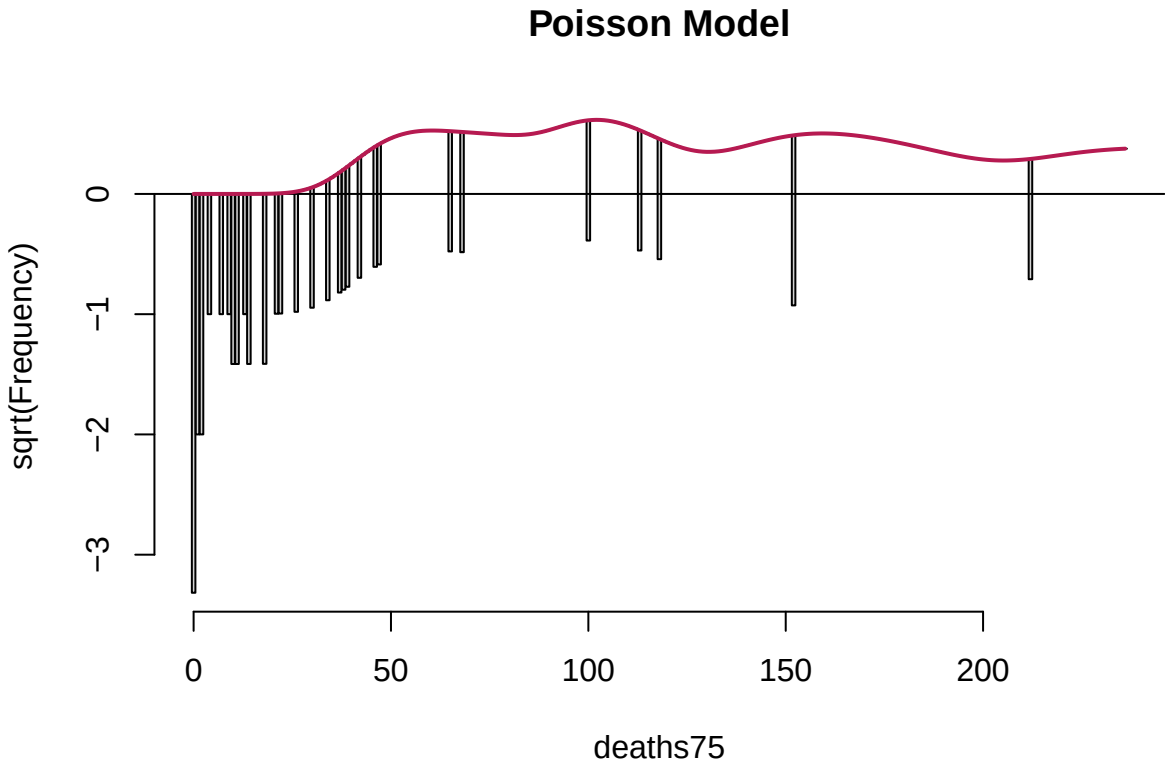
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, May 17, 2019 - 5:28:10 PM

Table 3:

	Dependent variable:
	deaths75
sanctions	0.005***
	(0.0001)
giniland	1.071***
	(0.053)
landless	0.045***
	(0.001)
energypc	−0.0001***
	(0.00001)
Constant	3.880***
	(0.041)
Observations	59
Log Likelihood	−21,191.300
Akaike Inf. Crit.	42,392.590
Note:	*p<0.1; **p<0.05; ***p<0.01

We get significance on all variables in the Poisson model, but we're still observing non-integer counts and a huge AIC (42,392).

```
rg1<-rootogram(p1, style = "hanging", scale = "sqrt", width = 1, plot = TRUE, main = "Poisson Model")
```




```
##rootogram(p1, main="Poisson", xlab="Mediations",
  ##ylim=c(-4,9),xlim=c(0,12), ylab=expression(sqrt(frequency)),
    ##pch = 19, col="black")
```

This rootogram indicates that overdispersion is present. Both the wave-like pattern and the severe underprediction of zeros indicate overdispersion. It is also underpredicting counts at the sample mean. Thin bar widths reflect the small n of this dataset.

```
d1 <- dispersiontest(p1,trafo=1)
print(d1) ## Please note: this function is not compatible with Stargazer.
```

```
##
##  Overdispersion test
##
## data:  p1
## z = 1.2604, p-value = 0.1038
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##    alpha
## 1987.357
```

Because our p-value is smaller than .05, we have no reason to reject the null hypothesis that the true alpha is greater than zero, implying overdispersion.

3. Fit the same model as in (2) but using a negative binomial specification. Discuss your results in a compelling fashion in no more that 1 page.

```
b1 <- glm.nb(deaths75 ~ sanctions + giniland + landless + energypc, data = ms)
```

```
stargazer(b1)
```

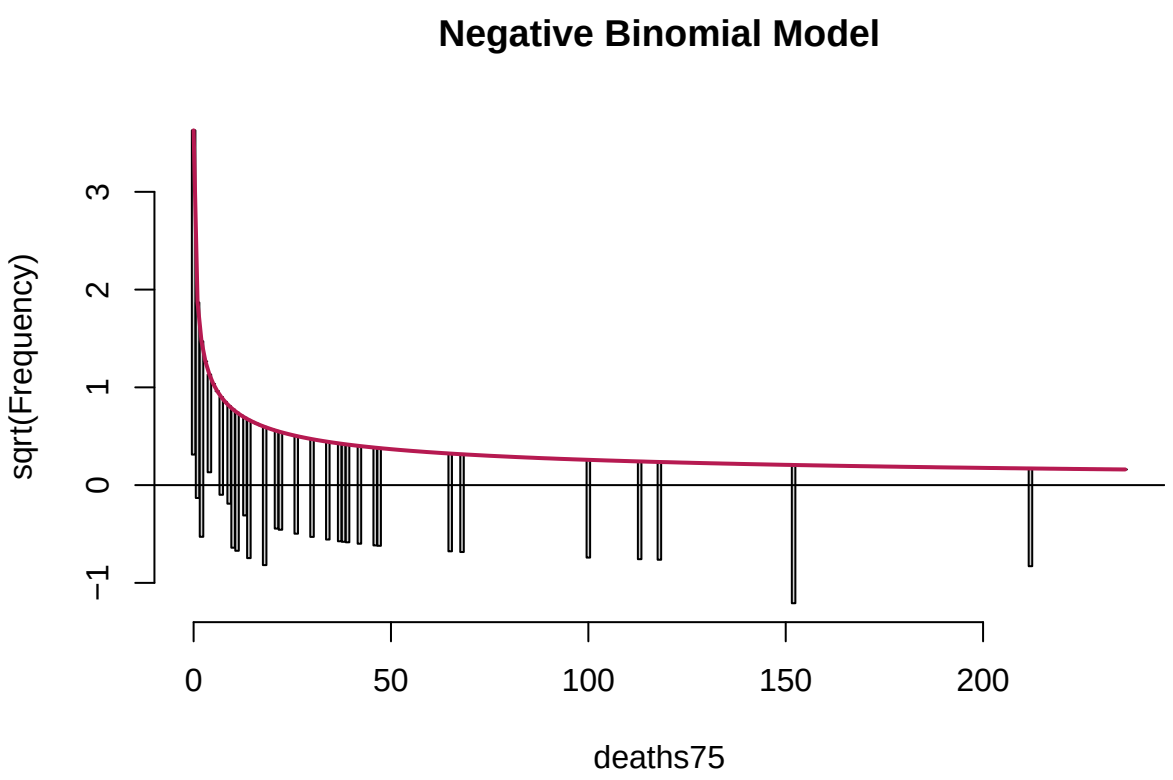
% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, May 17, 2019 - 5:28:11 PM

Table 4:

	Dependent variable:
	deaths75
sanctions	0.011^{***}
	(0.003)
giniland	3.825^{***}
	(1.421)
landless	0.050^{*}
	(0.029)
energypc	-0.0004^{***}
	(0.0001)
Constant	2.078^{**}
	(0.951)
Observations	59
Log Likelihood	-294.419
θ	0.273^{***} (0.046)
Akaike Inf. Crit.	598.838
Note:	*p<0.1; **p<0.05; ***p<0.01

These results indicate that we may be accounting for the overdispersion. We've retained the significance on all variables that we saw in the Poisson plot, but we've dramatically reduced AIC, and our value of theta is significant.

```
rg2<-rootogram(b1, style = "hanging", scale = "sqrt", width = 1, plot = TRUE, main = "Negative Binomial
```



This rootogram indicates that we are slightly underpredicting counts at most values, but the proximity to zero and the relatively small difference between observed and expected frequencies indicate that we have largely accounted for the overdispersion. Because the Poisson is a limiting case of the negative binomial, it should be possible to implement another over-dispersion test by constructing a likelihood ratio between the two models, per Ahlquist and Ward, Ch, 10, footnote #7.

```
odTest(b1)
```

```
## Likelihood ratio test of H0: Poisson, as restricted NB model:
## n.b., the distribution of the test-statistic under H0 is non-standard
## e.g., see help(odTest) for details/references
##
## Critical value of test statistic at the alpha= 0.05 level: 2.7055
## Chi-Square Test Statistic =  41795.7565 p-value = < 2.2e-16
```

We have strong reasons to reject the null hypothesis of the Poisson restriction in favor of our negative binomial model. This is because the test statistic (41,795) is significantly greater than 2.7055. Our p-value approaches zero, indicating confidence in this result.

```
stargazer(m1,m2,m3,m4,p1,b1)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Fri, May 17, 2019 - 5:28:11 PM

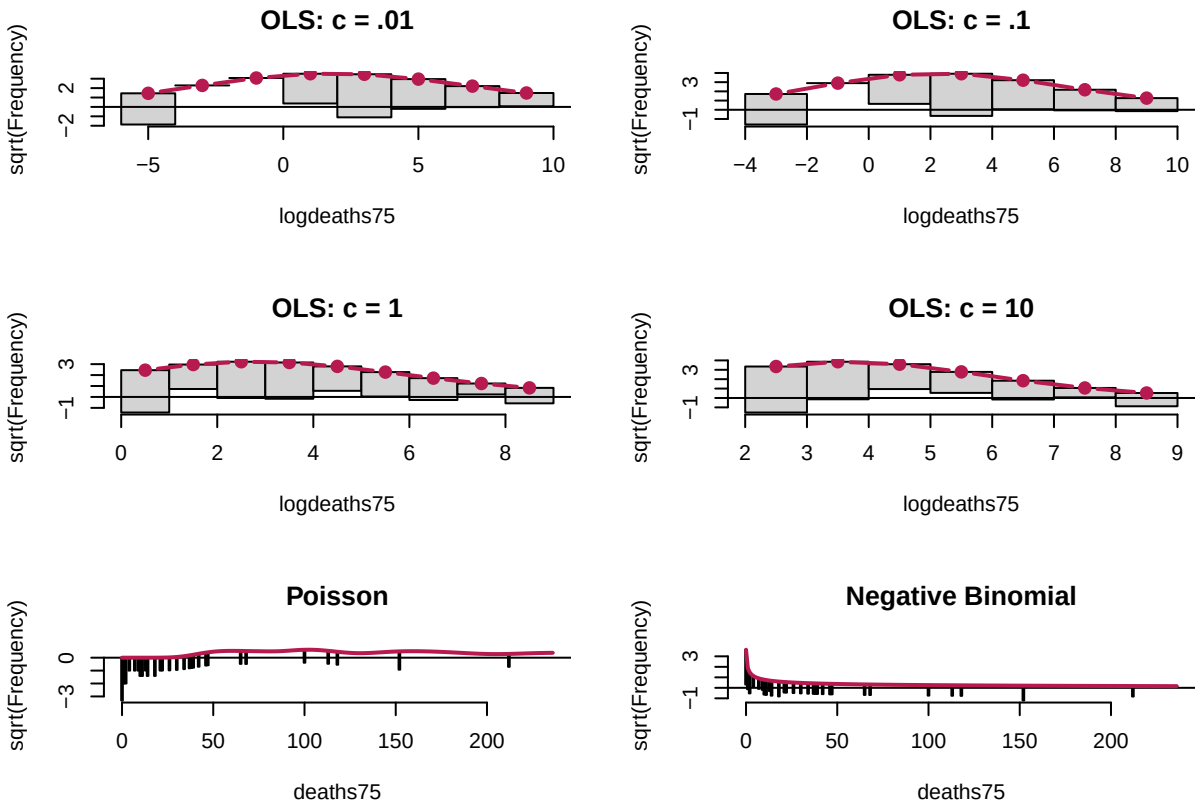
We observe a lower AIC and BIC for the OLS models, but OLS models normally have lower AIC and BIC (because of the homoskedasticity and normality assumptions). The AIC of the Poisson model is enormous, and the AIC for the OLS models is not directly comparable to the binomial AIC. These results give us strong reasons to prefer the negative binomial model.

Table 5:

	Dependent variable:					
	logdeaths75				deaths75	
	normal				*Poisson*	*negative binomial*
	(1)	(2)	(3)	(4)	(5)	(6)
sanctions	0.019***	0.016***	0.013***	0.009***	0.005***	0.011***
	(0.006)	(0.004)	(0.003)	(0.002)	(0.0001)	(0.003)
giniland	1.621	1.211	0.758	0.228	1.071***	3.825***
	(2.460)	(1.916)	(1.437)	(1.003)	(0.053)	(1.421)
landless	0.064	0.056	0.048	0.039*	0.045***	0.050*
	(0.051)	(0.039)	(0.030)	(0.021)	(0.001)	(0.029)
energypc	−0.0002	−0.0002	−0.0002	−0.0001	−0.0001***	−0.0004***
	(0.0002)	(0.0002)	(0.0001)	(0.0001)	(0.00001)	(0.0001)
Constant	−0.200	0.698	1.686*	2.999***	3.880***	2.078**
	(1.646)	(1.282)	(0.962)	(0.671)	(0.041)	(0.951)
Observations	59	59	59	59	59	59
Log Likelihood	−153.354	−138.603	−121.648	−100.444	−21,191.300	−294.419
θ						0.273*** (0.046)
Akaike Inf. Crit.	316.709	287.207	253.296	210.888	42,392.590	598.838

Note: *p<0.1; **p<0.05; ***p<0.01

```
par(mfrow=c(3,2))
rm1<-rootogram(m1, style = "hanging", scale = "sqrt", width = 1, plot = TRUE, main = 'OLS: c = .01')
rm2<-rootogram(m2, style = "hanging", scale = "sqrt", width = 1, plot = TRUE, main = 'OLS: c = .1')
rm3<-rootogram(m3, style = "hanging", scale = "sqrt", width = 1, plot = TRUE, main = 'OLS: c = 1')
rm4<-rootogram(m4, style = "hanging", scale = "sqrt", width = 1, plot = TRUE, main = 'OLS: c = 10')
rp1<-rootogram(p1, style = "hanging", scale = "sqrt", width = 1, plot = TRUE, main = 'Poisson')
rb1<-rootogram(b1, style = "hanging", scale = "sqrt", width = 1, plot = TRUE, main = 'Negative Binomial
```



Here I briefly show rootograms for all types of models. Though they are not directly comparable (because m1-m4 logged the 'deaths75' variable), we see that the binomial model is a more flexible distribution that is better able to capture heterogeneity in the rate parameter.)