Assignment 1

Michal Malyska

23/01/2020

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at /Users/michalmalyska/Desktop/University/Grad School/Classes/STA2201 - Applied Stati
```

```
library(aod)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
theme_set(theme_minimal())
```

Question 1

$$p(y|\theta, \phi) = exp\left(\frac{y\theta - b(\theta)}{\phi} - c(y, \theta)\right)$$

a)

Show $\int \frac{dp}{d\theta} dy = 0$ and $\int \frac{d^2p}{d\theta^2} dy = 0$

i)

Showing:

$$\int \frac{dp}{d\theta} dy = 0$$

$$
\begin{aligned}
\int \frac{dp}{d\theta}dy = \\
&= \frac{d}{d\theta}\int pdy \\
&= \frac{d}{d\theta}\int exp\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)\right)dy \\
&= \frac{d}{d\theta}(1) = 0
\end{aligned}
$$

ii)

Showing:

$$\int \frac{d^2p}{d\theta^2}dy = 0$$

$$
\begin{aligned}
\int \frac{dp}{d\theta}dy &= \frac{d^2}{d\theta^2}\int pdy \\
&= \frac{d^2}{d\theta^2}\int exp\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)\right)dy \\
&= \frac{d^2}{d\theta^2}(1) = 0
\end{aligned}
$$

b

i)

Showing $\mathbb{E}[Y] = b'(\theta)$

$$
\begin{aligned}
\frac{dp}{d\theta} &= \frac{d}{d\theta}\left(exp\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)\right)\right) \\
&= exp\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)\right) * \frac{d}{d\theta}\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\theta)\right) \\
&= p * \left(\frac{y}{\phi} - \frac{b'(\theta)}{\phi}\right)
\end{aligned}
$$

$$
\begin{aligned}
0 &= \int \frac{dp}{d\theta}dy \\
&= \int p * \left(\frac{y}{\phi} - \frac{b'(\theta)}{\phi}\right)dy \\
&= \frac{1}{\phi}(\mathbb{E}[Y] - b'(\theta)) \\
&\implies \mathbb{E}[Y] = b'(\theta)
\end{aligned}
$$

ii)

Showing $\mathbb{V}ar(Y) = \phi b''(\theta)$

$$\begin{aligned}
\frac{d^2p}{d\theta^2} &= \frac{d^2}{d\theta^2}\left(exp\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)\right)\right) \\
&= \frac{d}{d\theta}\left(exp\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)\right) * \frac{d}{d\theta}\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\theta)\right)\right) \\
&= \frac{d}{d\theta}\left(p * \left(\frac{y}{\phi} - \frac{b'(\theta)}{\phi}\right)\right) \\
&= \frac{dp}{d\theta}\left(\frac{y}{\phi} - \frac{b'(\theta)}{\phi}\right) - p * \left(\frac{b''(\theta)}{\phi}\right)
\end{aligned}$$

$$\begin{aligned}
0 &= \int \frac{d^2p}{d\theta^2} dy \\
&= \int \frac{dp}{d\theta}\left(\frac{y}{\phi} - \frac{b'(\theta)}{\phi}\right) - p * \left(\frac{b''(\theta)}{\phi}\right) dy \\
&= \int \frac{dp}{d\theta}\left(\frac{y}{\phi} - \frac{b'(\theta)}{\phi}\right) dy - \int p * \left(\frac{b''(\theta)}{\phi}\right) dy \\
&= \int p * \left(\frac{y}{\phi} - \frac{b'(\theta)}{\phi}\right)^2 dy - \frac{b''(\theta)}{\phi} \\
&= \frac{1}{\phi^2}\left(\mathbb{V}ar[Y] + 0 - \phi b''(\theta)\right) \\
&\implies \mathbb{V}ar[Y] = \phi b''(\theta)
\end{aligned}$$

c

i)

Showing that $\mathbb{E}[\frac{dl}{d\theta}] = 0$

I will denote $l = l(\theta)$ for simplicity

$$\begin{aligned}
\mathbb{E}[\frac{dl}{d\theta}] &= \mathbb{E}\left[\frac{d}{d\theta}\left(\frac{y\theta - b(\theta)}{\phi} - c(y,\phi)\right)\right] \\
&= \mathbb{E}\left[\frac{y - b'(\theta)}{\phi}\right] \\
&= \frac{1}{\phi}(\mathbb{E}\left[y\right] - b'(\theta)) = 0
\end{aligned}$$

ii)

Showing that $\mathbb{V}ar[\frac{dl}{d\theta}] = \phi^{-1}b''(\theta)$

$$\begin{aligned}
\mathbb{V}ar\left[\frac{dl}{d\theta}\right] &= \mathbb{E}\left[\left(\frac{dl}{d\theta}\right)^2\right] \\
&= -\mathbb{E}\left[\frac{d^2l}{d\theta^2}\right] \\
&= -\mathbb{E}\left[\frac{d}{d\theta}\left(\frac{y - b'(\theta)}{\phi}\right)\right] \\
&= \mathbb{E}\left[\left(\frac{b''(\theta)}{\phi}\right)\right] \\
&= \frac{b''(\theta)}{\phi}
\end{aligned}$$

Question 2

a

$Y|\theta \sim Poisson(\mu\theta)$

$\mathbb{E}[\theta] = 1$ and $\mathbb{V}ar[\theta] = \sigma^2$

i)

Showing $\mathbb{E}[Y] = \mu$

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}\left[\mathbb{E}[Y|\theta]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{e^{-\mu\theta}(\mu\theta)^y}{y!}\right]\right] \\
&= \mathbb{E}\left[\mu\theta\right] \\
&= \mu
\end{aligned}$$

ii)

Showing $\mathbb{V}ar[Y] = \mu(1 + \mu\sigma^2)$

$$\begin{aligned}
\mathbb{V}ar[Y] &= \mathbb{E}\left[\mathbb{V}ar(Y|\theta)\right] + \mathbb{V}ar\left[\mathbb{E}(Y|\theta)\right] \\
&= \mathbb{E}\left[\mu\theta\right] + \mathbb{V}ar\left[\mu\theta\right] \\
&= \mu + \mu^2\sigma^2 \\
&= \mu(1 + \mu\sigma^2)
\end{aligned}$$

b

Assume $\theta \sim \Gamma(\alpha, \beta)$

Showing $Y \sim NegBin$

$$\begin{aligned}
p(y) &= \int p(y|\theta)p(\theta)d\theta \\
&= \int \frac{e^{-\mu\theta}(\mu\theta)^y}{y!} * \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}d\theta \\
&= \frac{\mu^y}{\beta^\alpha\Gamma(\alpha)y!}\int e^{-\mu\theta}\theta^y\theta^{\alpha-1}e^{-\theta/\beta}d\theta \\
&= \frac{\mu^y}{\beta^\alpha\Gamma(\alpha)y!}\int e^{-(\mu+1/\beta)\theta}\theta^{y+\alpha-1}d\theta \\
&= \frac{\mu^y}{\beta^\alpha\Gamma(\alpha)y!} * \left(\Gamma(y+\alpha)(\frac{\beta}{\beta\mu+1})^{\alpha+y}\right) \\
&= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)} * \frac{\mu^y\beta^{\alpha+y}}{\beta^\alpha} * (\beta\mu+1)^{-\alpha-y} \\
&= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)}\left(\frac{\mu\beta}{\mu\beta+1}\right)^y\left(\frac{1}{\mu\beta+1}\right)^\alpha
\end{aligned}$$

$$
\begin{aligned}
p(y) &= \int p(y|\theta)p(\theta)d\theta \\
&= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)}\left(\frac{\mu\beta}{\mu\beta+1}\right)^y\left(\frac{1}{\mu\beta+1}\right)^\alpha \\
&= NB(\alpha, \frac{\mu\beta}{\mu\beta+1})
\end{aligned}
$$

c

$$
\begin{aligned}
\mathbb{E}[Y] = \mu = \alpha\mu\beta &\implies \alpha\beta = 1 \\
\mathbb{V}ar[Y] = \mu + \mu^2\sigma^2 = \alpha\mu\beta + \alpha\mu^2\beta^2 &\implies \alpha\beta^2 = \sigma^2
\end{aligned}
$$

$$
\begin{aligned}
\alpha &= \frac{1}{\sigma^2} \\
\beta &= \sigma^2
\end{aligned}
$$

Question 3

I refactored the code a tiny bit

```
set.seed(123)

X <- matrix(NA, 100, 100)
Y <- X
for (i in 1:100) {
  X[i, ] <- rnorm(100)
  Y[i, ] <- rpois(100, lambda = exp(0.5 + X[i, ] + 0.2 * X[i, ]^2))
}
```

a) Fitting poisson glm

```
coefs_matrix <- matrix(NA, 100, 3)
ses_matrix <- matrix(NA, 100, 3)
p_vals_check <- rep(NA, 100)

for (i in 1:100) {
  data_set <- tibble(x = X[i, ], y = Y[i, ])
  mod <-
    glm(
      formula = y ~ x + I(x^2),
      data = data_set,
      family = poisson
    )
  coefs_matrix[i, ] <- coefficients(mod)
  ses_matrix[i, ] <- sqrt(diag(vcov(mod)))
  p_vals_check[i] <-
    wald.test(
      b = coef(mod),
      Sigma = vcov(mod),
      Terms = 2,
      H0 = 1
    )$result$chi2[3]
}
```

b) coverage probability for 2SE on x

Since this is an MLE blah blah blah it's enough to look at normal CDF up to 2 sd so the coverage is 0.9772499

The actual proportion of coefficients outside of the intervals is 4 which is 4% for a coverage probability of ~96%

Is this valid for x? Not 100% since the variables are not independent, in principle they should be uncorrelated but in practice their cor is 0.2974821 this will definitely fudge with inference, but hopefully in a minor way.

Also, doesn't really match the 95% CI thing.

c) Wald tests

```
p_vals <- rep(NA, 100)

for (i in 1:100) {
  W <- (coefs_matrix[i, 2] - 1) / ses_matrix[i, 2]
  p_vals[i] <- 1 - (pnorm(abs(W)) - pnorm(-abs(W)))
}
```

Test was rejected in 4 case(s).

```
set.seed(321)

X2 <- matrix(NA, 100, 100)
Y2 <- X2
for (i in 1:100) {
  weights <- ifelse(X[i, ] > 1, 10, 1)
  probs <- weights / sum(weights)
  to_keep_2 <- sample(1:length(X[i, ]), 25, prob = probs)
  X2[i, ] <- X[i, to_keep_2]
  Y2[i, ] <- Y[i, to_keep_2]
}
```

d)

GLMs

```
coefs_matrix2 <- matrix(NA, 100, 3)
ses_matrix2 <- matrix(NA, 100, 3)


for (i in 1:100) {
  data_set <- tibble(x = X2[i, ], y = Y2[i, ])
  mod <- glm(
    formula = y ~ x + I(x^2),
    data = data_set,
    family = poisson
  )
  coefs_matrix2[i, ] <- coefficients(mod)
  ses_matrix2[i, ] <- sqrt(diag(vcov(mod)))
}

num_inside_interval <-
  sum(coefs_matrix2[, 2] + 2 * ses_matrix2[, 2] < 1) + sum(coefs_matrix2[, 2] - 2 * ses_matrix2[, 2] >
```

Coverage probabilities:

The actual proportion of coefficients outside of the intervals is 28 which is 28% for a coverage probability of 72%

Wald tests:

```r
p_vals2 <- rep(NA, 100)


for (i in 1:100) {
  W <- (coefs_matrix2[i, 2] - 1) / ses_matrix2[i, 2]
  p_vals2[i] <- 1 - (pnorm(abs(W)) - pnorm(-abs(W)))
}
```

Test was rejected in 29 cases which more or less agrees with the coverages calculated before.

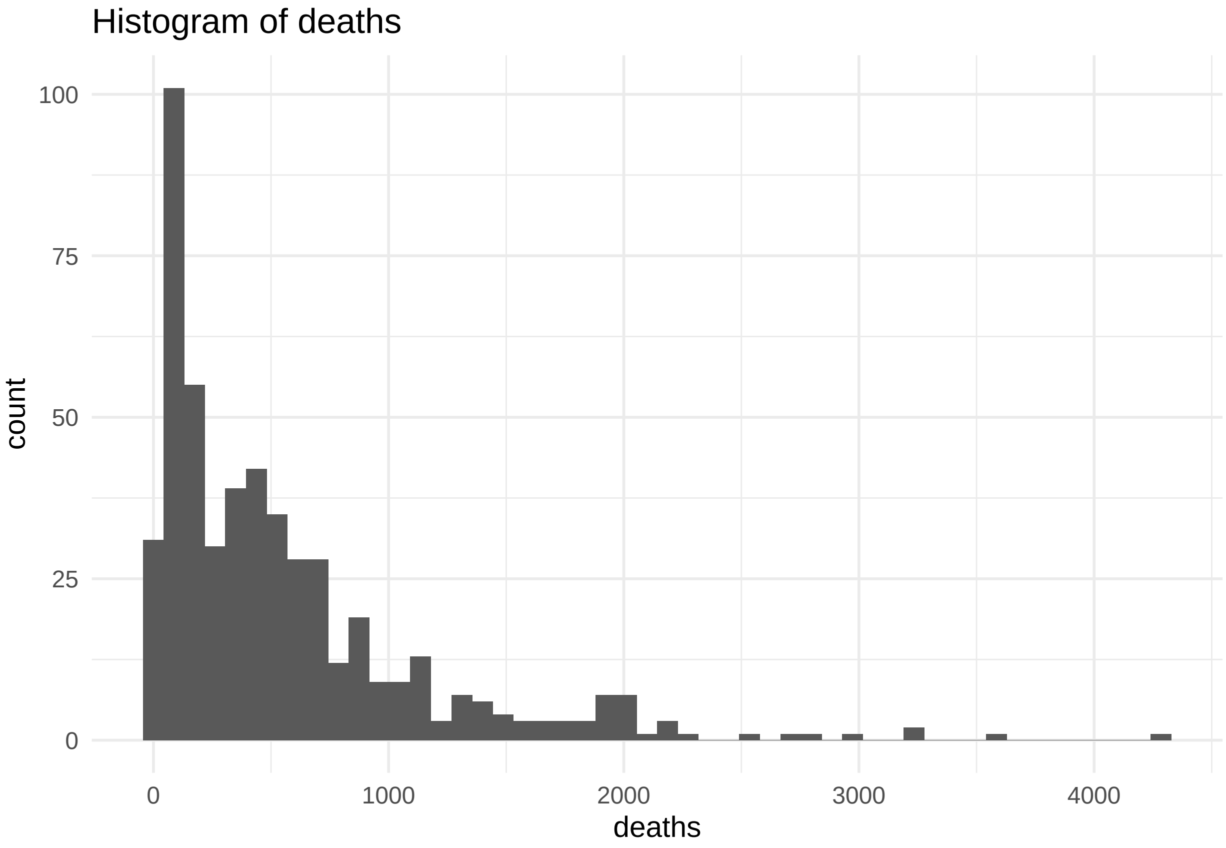
e)

What's happening is that we now have some selection process in the data. In this particular case, high values of x were more likely to show up in the dataset.

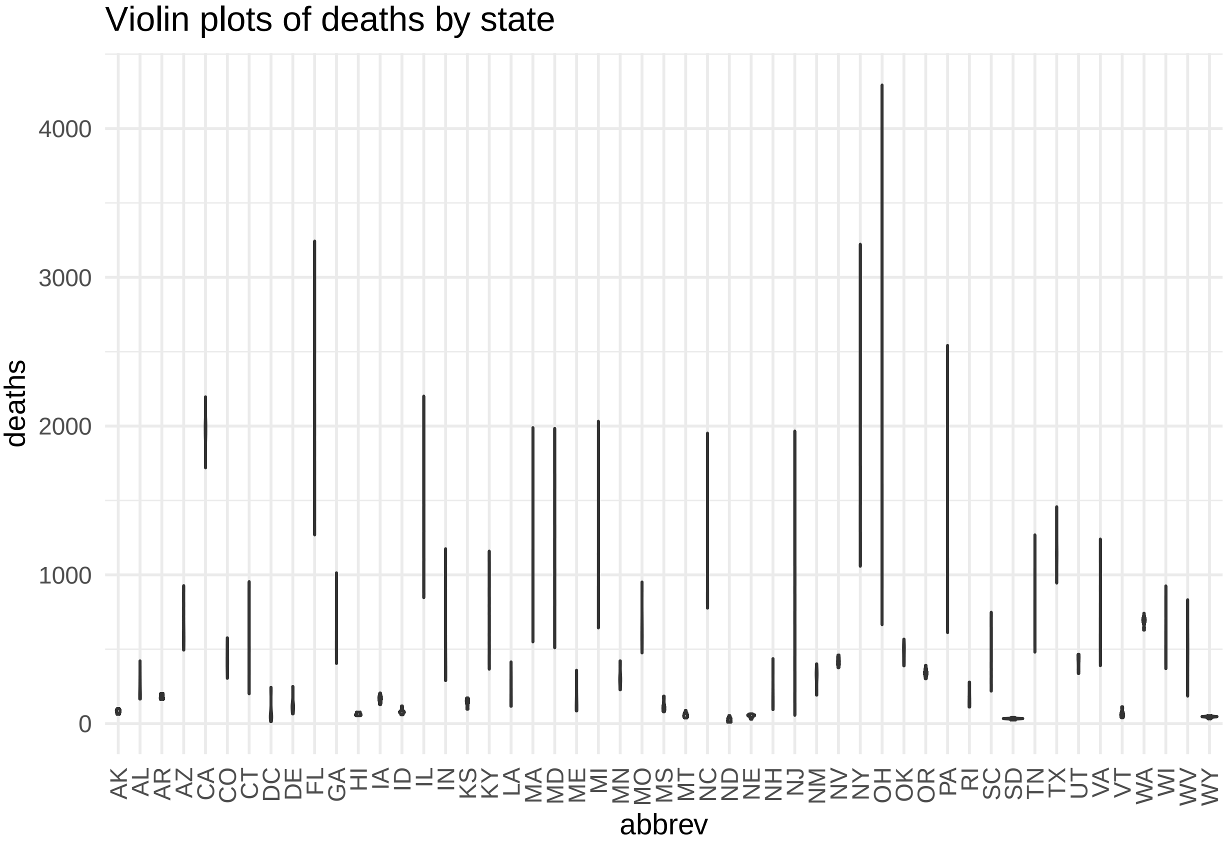
Question 4

a) - EDA

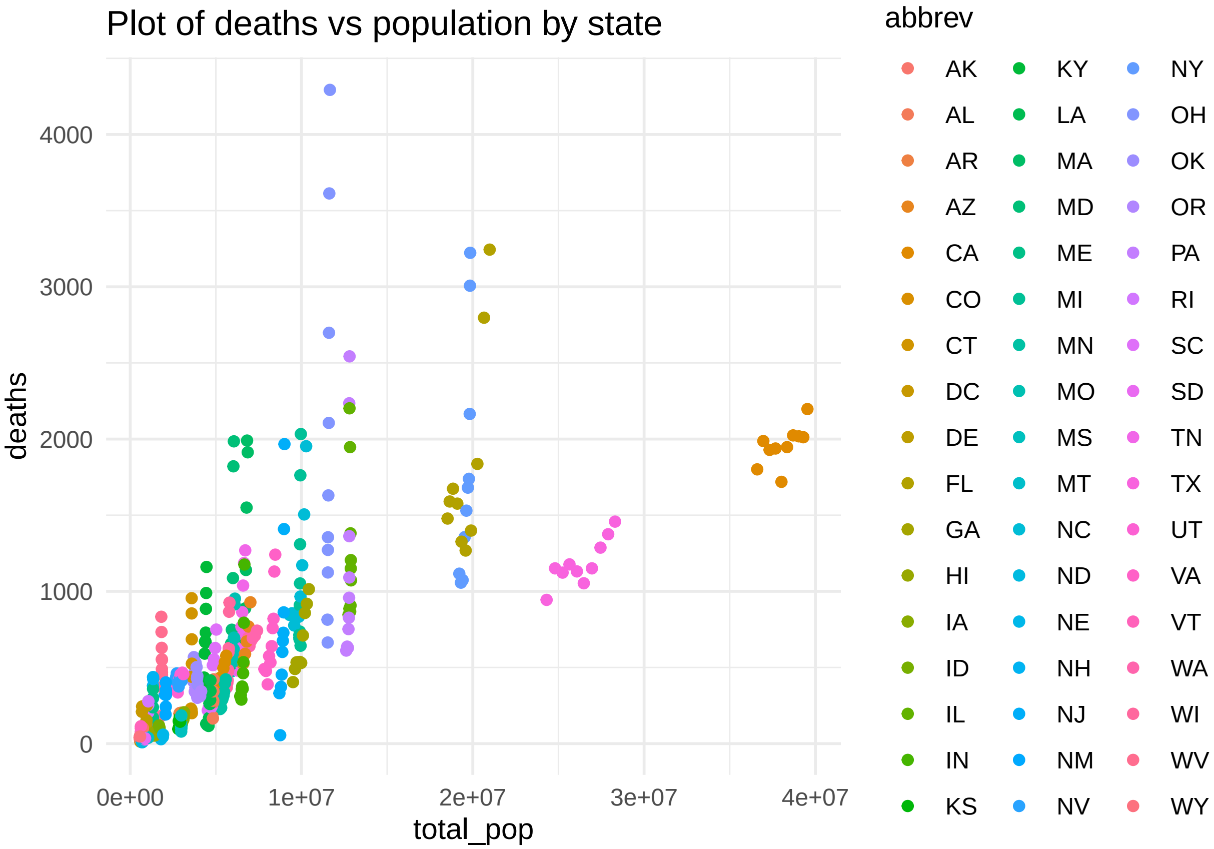
First I will generate a ton of plots of variables to visually look for patterns.



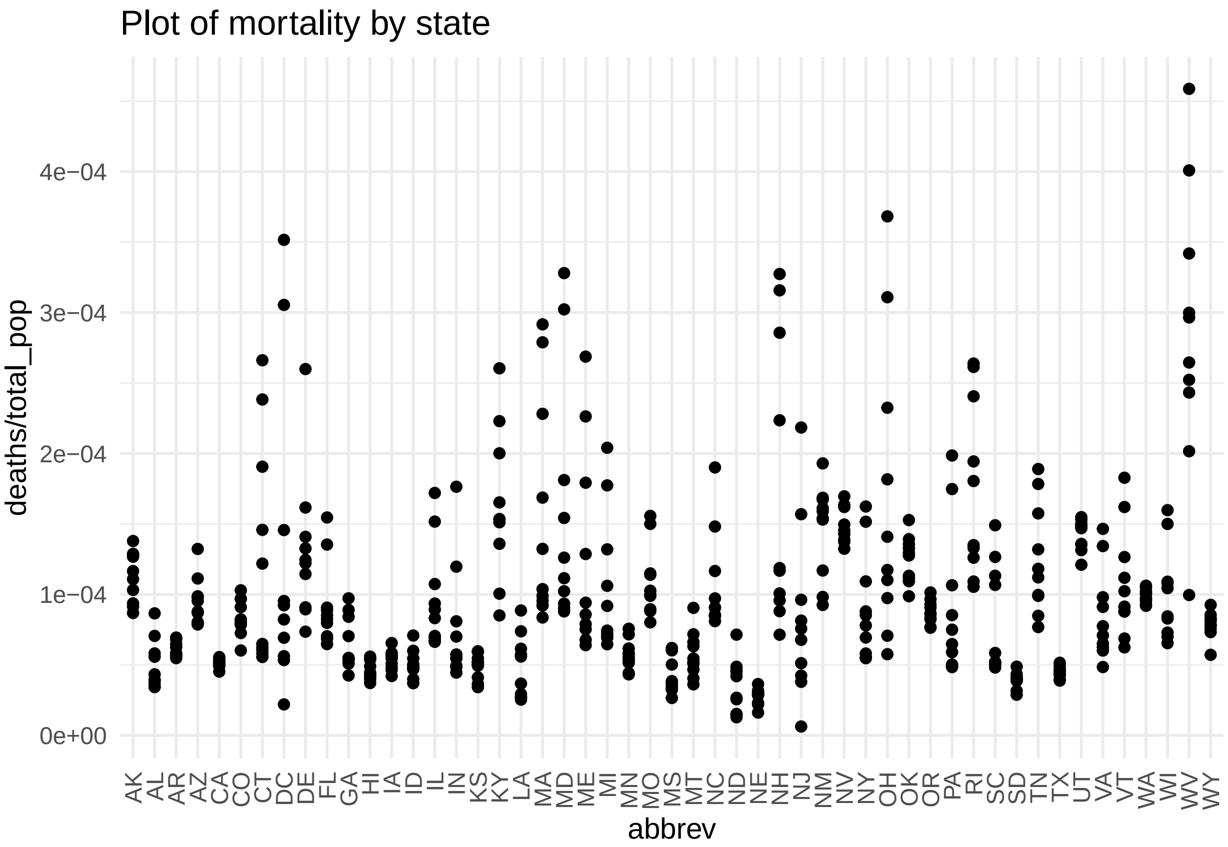
I can tell that the distribution is right-skewed and with quite some observations out in the high deathcounts. This is without context so next I wanna see if some particular states have large variations (of course they do)



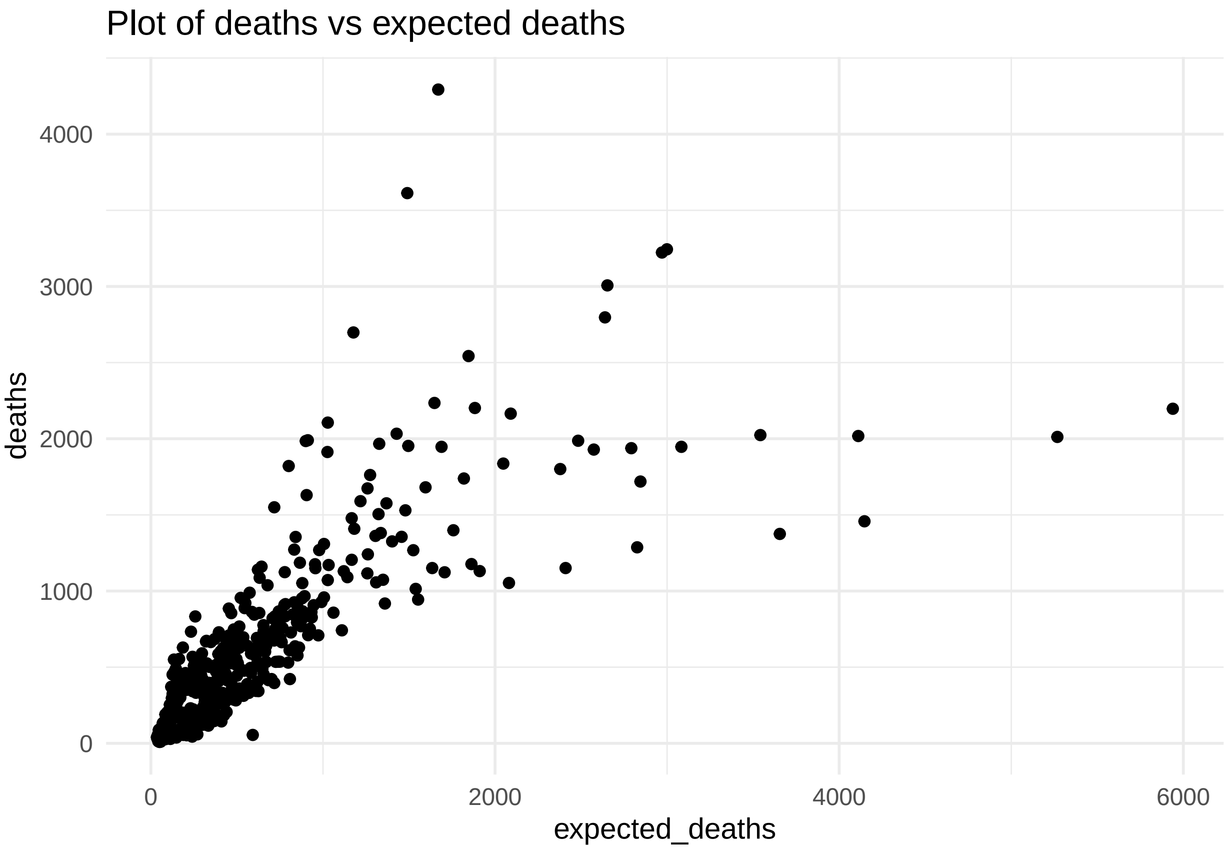
There are a couple of states with huge variations like OH (Ohio?) or FL (Florida) and quite a few of the states have very low variations and low numbers. This is not as likely due to just population since California would be somewhere in the sky. Next I'm gonna make sure that it's true by looking at deaths vs pop and color the states.



There seems to be a population pattern to some degree (fair), but overall there seems to be quite a lot of variation that's outside of that. Let's look at mortality (deaths / pop) to see if there is something a bit easier to spot.

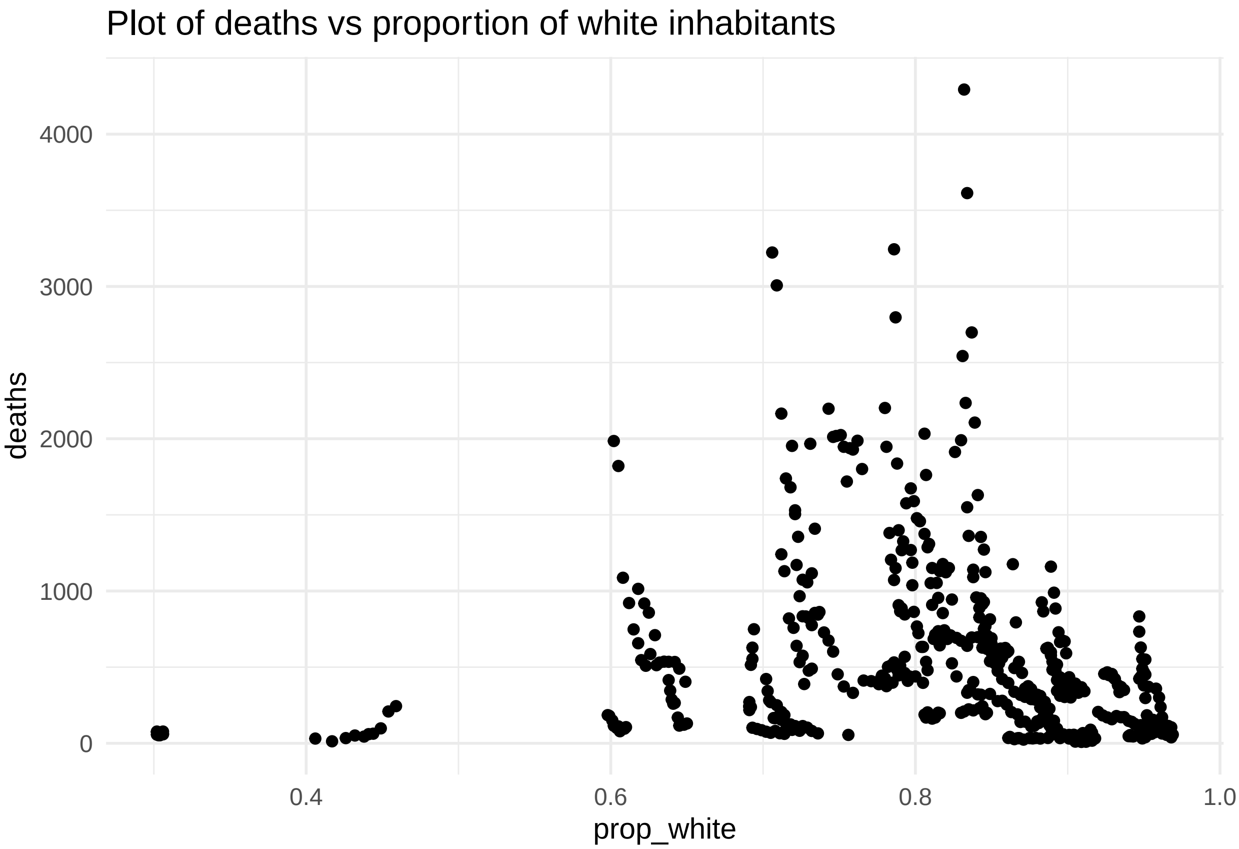


This should show the variation in deaths that are not exactly just due to high pop. Clearly there are some states that are way out there (again OH). Let's check out the expected deaths vs actual

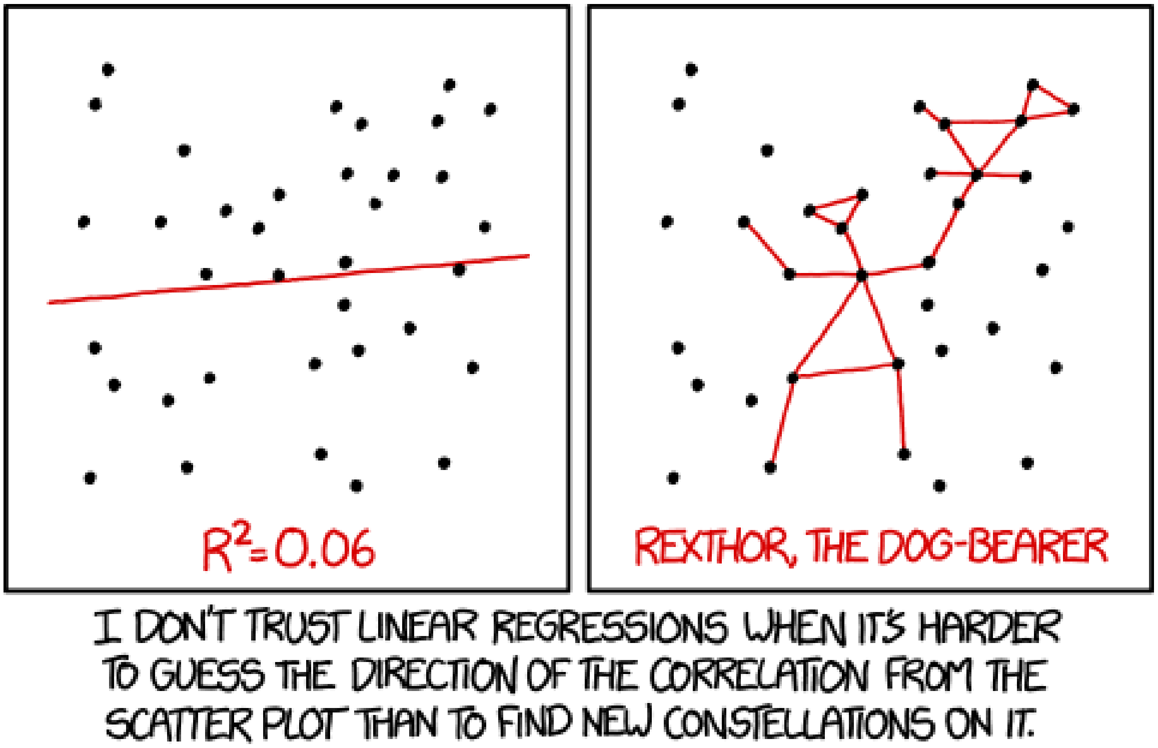


This seems to make a lot better at predicting the actual deaths. Probably a solid variable to use but by the data dictionary provided it's a derivative of the other variables provided so probably unwise to use it alongside them and make statements about those variables' coefficients.

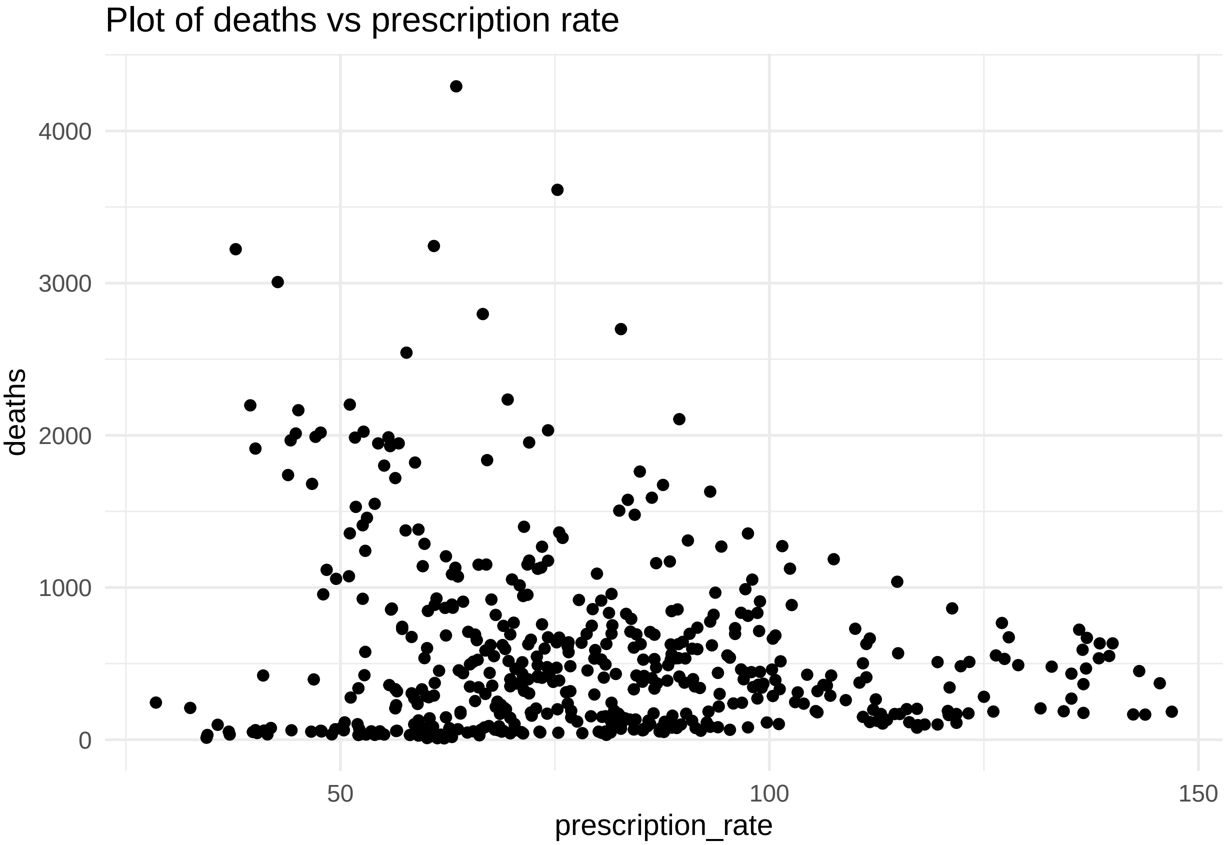
Let's check out whiteness



I can see a woman yelling at a worm. This is a typical example of this:

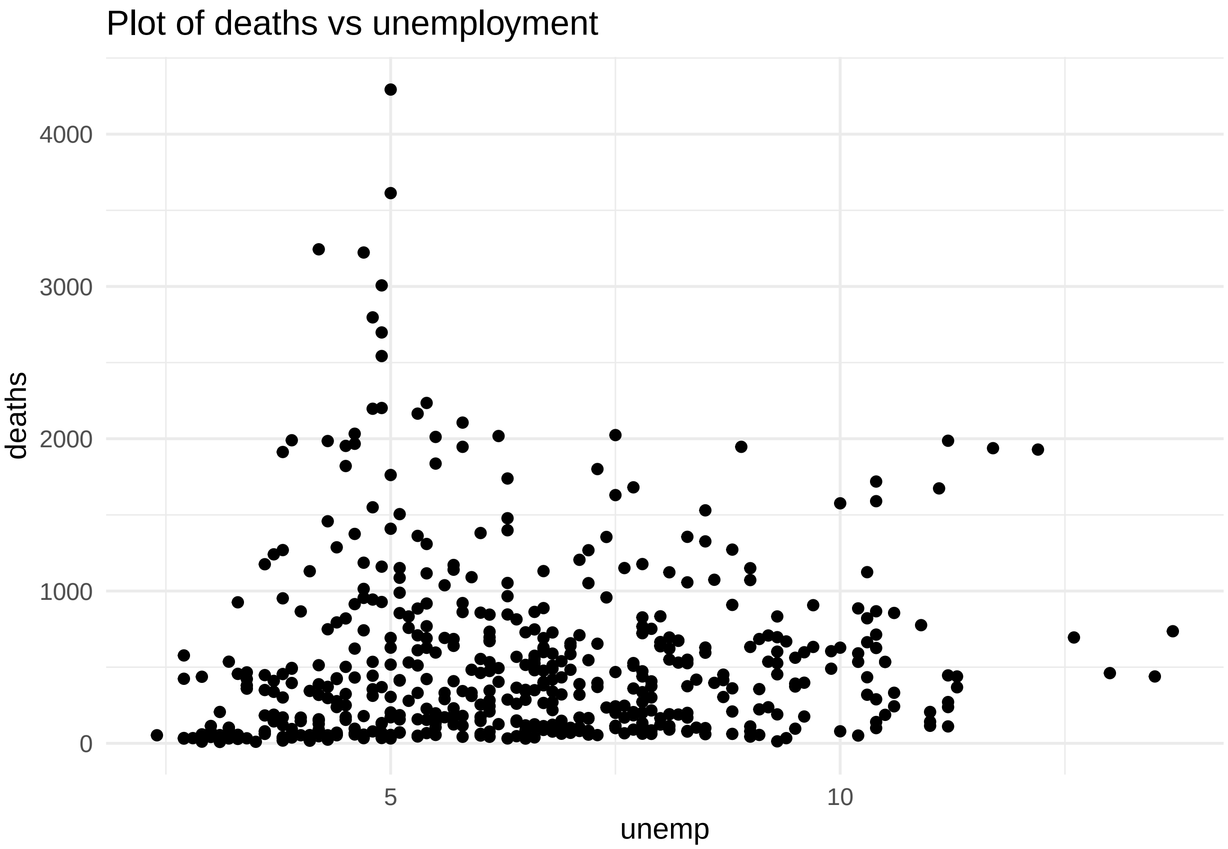

R²=0.06
REXTHOR, THE DOG-BEARER
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

I don't think it's worth using it as a predictor. I could probably massage this a bit and get something for the model but I highly doubt it would have any real meaning. Let's check out prescription rates:

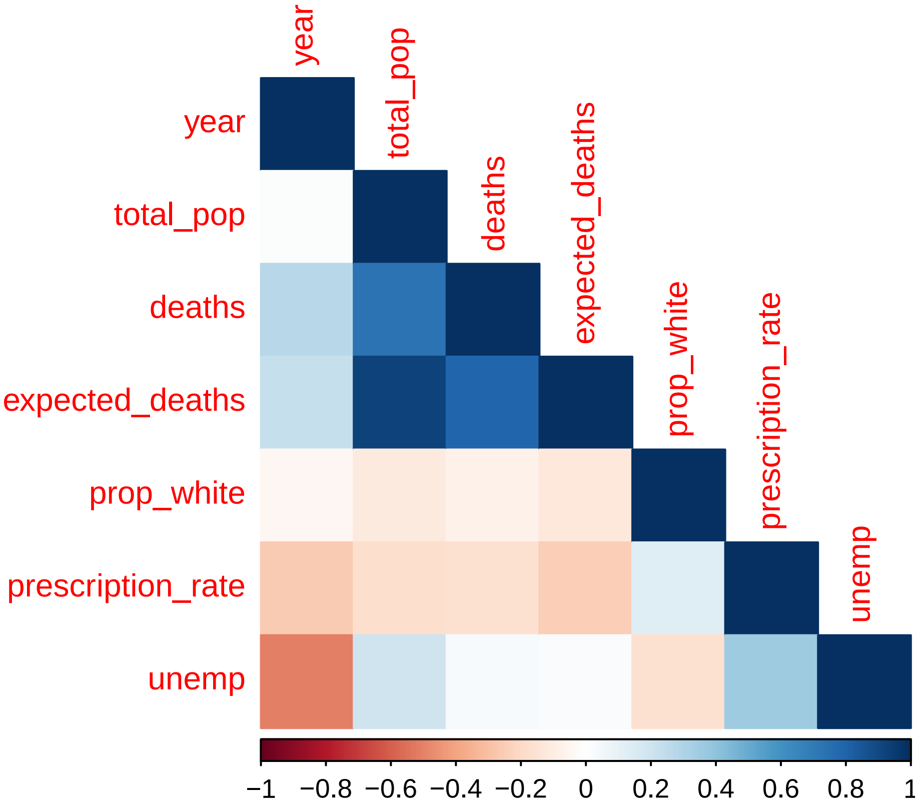


Again there seems to be very minimal trend with overall prescription rates. I don't think it's that good of a var. I could try to massage it a bit by getting it to be prescription numbers

Finally let's take a look at the situation in the job market:



Again there doesn't seem to be that much of a pattern. I don't know if the variable is worthwhile to use. Let's take a peek at the correlations



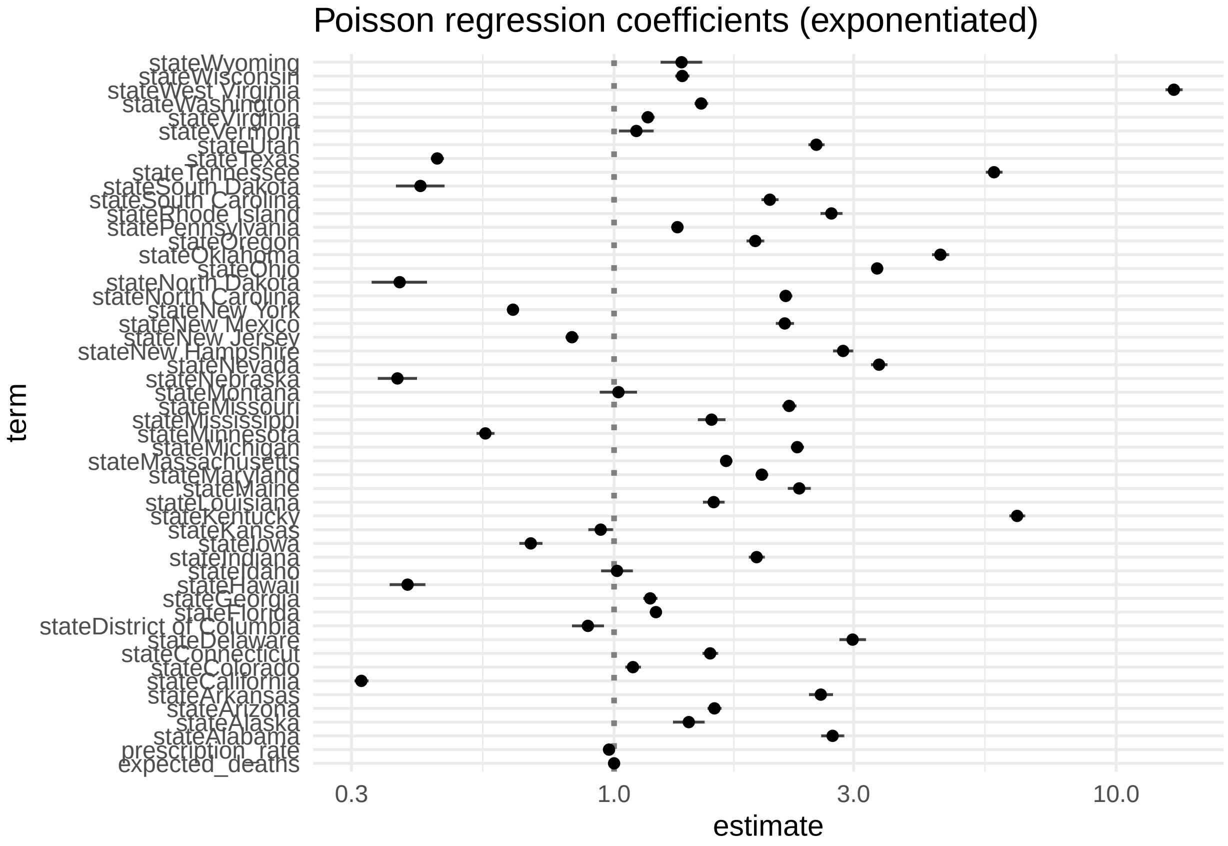
I think the predictors to include overall are (either expected_deaths or total_pop), plus state and prescription rate.

Let’s look at some summaries:

state	n	mean_deaths	var_deaths	median_deaths	min_deaths	max_deaths	mean_mort
Illinois	10	1246.3	2.236551e+05	1111.0	846	2202	0.0000971
Alabama	10	240.2	7.620400e+03	196.5	165	422	0.0000498
Alaska	10	81.4	1.687111e+02	83.5	62	102	0.0001124
Arizona	10	629.7	1.740801e+04	597.0	494	928	0.0000949
Arkansas	10	182.0	2.333333e+02	180.5	162	203	0.0000618
California	10	1957.2	1.688440e+04	1967.0	1719	2197	0.0000513
Colorado	10	443.9	7.330989e+03	425.5	304	577	0.0000843
Connecticut	10	453.7	8.390557e+04	334.5	200	955	0.0001264
Delaware	10	121.8	2.731733e+03	113.0	65	250	0.0001310
District of Columbia	10	84.7	6.163567e+03	55.5	13	244	0.0001273
Florida	10	1818.9	4.399199e+05	1583.0	1268	3244	0.0000921
Georgia	10	652.9	4.347699e+04	535.5	404	1014	0.0000651
Hawaii	10	64.3	8.734444e+01	62.0	53	77	0.0000462
Idaho	10	81.6	3.307111e+02	77.5	59	119	0.0000504
Indiana	10	497.7	7.944357e+04	367.5	289	1176	0.0000756
Iowa	10	165.4	5.224889e+02	170.5	127	206	0.0000536
Kansas	10	140.8	6.999556e+02	147.0	96	173	0.0000489
Kentucky	10	716.0	5.837822e+04	671.0	365	1160	0.0001628
Louisiana	10	223.3	1.138179e+04	214.5	116	415	0.0000483
Maine	10	168.8	9.489289e+03	119.5	85	359	0.0001269
Maryland	10	937.3	2.950753e+05	702.5	509	1985	0.0001577
Massachusetts	10	1059.3	3.153162e+05	789.5	549	1990	0.0001574
Michigan	10	1053.8	2.440495e+05	822.5	643	2033	0.0001062
Minnesota	10	308.9	3.970544e+03	297.5	227	422	0.0000570
Mississippi	10	122.2	1.337289e+03	108.5	79	185	0.0000410
Missouri	10	660.7	2.539846e+04	609.0	475	952	0.0001094
Montana	10	57.5	2.282778e+02	53.5	38	89	0.0000572
Nebraska	10	51.4	1.131556e+02	54.0	29	66	0.0000276
Nevada	10	419.4	7.024889e+02	415.5	375	461	0.0001504
New Hampshire	10	231.8	1.891818e+04	155.5	94	437	0.0001744
New Jersey	10	745.5	3.160721e+05	638.5	55	1967	0.0000834
New Mexico	10	303.2	4.759511e+03	322.5	191	402	0.0001464
New York	10	1794.8	6.055146e+05	1605.5	1057	3223	0.0000912
North Carolina	10	1056.0	1.494104e+05	850.5	776	1953	0.0001069
North Dakota	10	25.1	2.107667e+02	24.5	9	54	0.0000348
Ohio	10	1956.9	1.486807e+06	1492.5	664	4293	0.0001687
Oklahoma	10	477.4	3.344489e+03	492.5	388	568	0.0001251
Oregon	10	342.3	7.051222e+02	341.0	301	392	0.0000872
Pennsylvania	10	1164.5	4.765249e+05	892.5	611	2543	0.0000912
Rhode Island	10	184.6	4.382044e+03	166.0	111	279	0.0001749
South Carolina	10	390.0	4.004022e+04	259.0	218	749	0.0000808
South Dakota	10	32.7	2.356667e+01	33.5	24	42	0.0000391
Tennessee	10	812.7	7.370868e+04	745.0	480	1269	0.0001247
Texas	10	1185.0	2.285267e+04	1151.0	944	1458	0.0000449
Utah	10	411.6	2.315600e+03	427.0	336	466	0.0001426
Vermont	10	67.7	5.855667e+02	60.5	39	114	0.0001084
Virginia	10	705.3	8.141423e+04	607.5	389	1241	0.0000855
Washington	10	687.4	1.114711e+03	693.5	628	742	0.0000989
West Virginia	10	526.3	3.337112e+04	520.0	184	833	0.0002859
Wisconsin	10	577.7	3.687623e+04	541.0	369	926	0.0001007
Wyoming	10	45.2	3.573333e+01	46.5	32	54	0.0000788

b) Poisson Regression

```
##
## Call:
## glm(formula = deaths ~ expected_deaths + state + prescription_rate,
##     family = poisson, data = df, offset = log(total_pop))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -30.6815   -3.1970   -0.0585    2.7924   24.5650
##
## Coefficients:
##                             Estimate Std. Error  z value Pr(>|z|)
## (Intercept)               -7.982e+00  1.993e-02 -400.423  < 2e-16 ***
## expected_deaths            1.186e-04  4.565e-06   25.980  < 2e-16 ***
## stateAlabama               1.002e+00  2.634e-02   38.053  < 2e-16 ***
## stateAlaska                3.427e-01  3.650e-02    9.389  < 2e-16 ***
## stateArizona               4.603e-01  1.581e-02   29.120  < 2e-16 ***
## stateArkansas              9.480e-01  2.746e-02   34.519  < 2e-16 ***
## stateCalifornia           -1.159e+00  1.591e-02  -72.850  < 2e-16 ***
## stateColorado              8.653e-02  1.766e-02    4.899 9.65e-07 ***
## stateConnecticut           4.406e-01  1.771e-02   24.883  < 2e-16 ***
## stateDelaware              1.094e+00  3.051e-02   35.856  < 2e-16 ***
## stateDistrict of Columbia -1.200e-01  3.678e-02   -3.261  0.00111 **
## stateFlorida               1.920e-01  1.290e-02   14.886  < 2e-16 ***
## stateGeorgia               1.657e-01  1.600e-02   10.354  < 2e-16 ***
## stateHawaii               -9.469e-01  4.110e-02  -23.040  < 2e-16 ***
## stateIdaho                 1.346e-02  3.645e-02    0.369  0.71197
## stateIndiana               6.538e-01  1.812e-02   36.071  < 2e-16 ***
## stateIowa                 -3.820e-01  2.644e-02  -14.447  < 2e-16 ***
## stateKansas               -6.140e-02  2.846e-02   -2.157  0.03099 *
## stateKentucky              1.848e+00  1.794e-02  103.039  < 2e-16 ***
## stateLouisiana             4.569e-01  2.470e-02   18.501  < 2e-16 ***
## stateMaine                 8.490e-01  2.628e-02   32.304  < 2e-16 ***
## stateMaryland              6.775e-01  1.389e-02   48.790  < 2e-16 ***
## stateMassachusetts         5.140e-01  1.352e-02   38.018  < 2e-16 ***
## stateMichigan              8.400e-01  1.473e-02   57.031  < 2e-16 ***
## stateMinnesota            -5.901e-01  2.053e-02  -28.744  < 2e-16 ***
## stateMississippi           4.470e-01  3.165e-02   14.122  < 2e-16 ***
## stateMissouri              8.029e-01  1.602e-02   50.135  < 2e-16 ***
## stateMontana               2.001e-02  4.287e-02    0.467  0.64076
## stateNebraska             -9.933e-01  4.519e-02  -21.980  < 2e-16 ***
## stateNevada                1.215e+00  1.866e-02   65.097  < 2e-16 ***
## stateNew Hampshire         1.050e+00  2.297e-02   45.711  < 2e-16 ***
## stateNew Jersey           -1.931e-01  1.482e-02  -13.029  < 2e-16 ***
## stateNew Mexico            7.827e-01  2.060e-02   37.992  < 2e-16 ***
## stateNew York             -4.635e-01  1.204e-02  -38.503  < 2e-16 ***
## stateNorth Carolina        7.870e-01  1.447e-02   54.375  < 2e-16 ***
## stateNorth Dakota         -9.832e-01  6.402e-02  -15.356  < 2e-16 ***
## stateOhio                  1.206e+00  1.285e-02   93.851  < 2e-16 ***
## stateOklahoma              1.496e+00  1.953e-02   76.630  < 2e-16 ***
## stateOregon                6.475e-01  1.998e-02   32.414  < 2e-16 ***
## statePennsylvania          2.906e-01  1.333e-02   21.796  < 2e-16 ***
## stateRhode Island          9.966e-01  2.528e-02   39.427  < 2e-16 ***
## stateSouth Carolina        7.143e-01  1.951e-02   36.620  < 2e-16 ***
## stateSouth Dakota         -8.878e-01  5.633e-02  -15.761  < 2e-16 ***
## stateTennessee             1.742e+00  1.893e-02   92.025  < 2e-16 ***
## stateTexas                -8.110e-01  1.477e-02  -54.915  < 2e-16 ***
## stateUtah                  9.276e-01  1.840e-02   50.417  < 2e-16 ***
## stateVermont               1.022e-01  3.994e-02    2.560  0.01048 *
## stateVirginia              1.551e-01  1.501e-02   10.329  < 2e-16 ***
## stateWashington            3.995e-01  1.524e-02   26.221  < 2e-16 ***
## stateWest Virginia         2.567e+00  1.948e-02  131.762  < 2e-16 ***
## stateWisconsin             3.126e-01  1.611e-02   19.407  < 2e-16 ***
## stateWyoming               3.095e-01  4.808e-02    6.437 1.22e-10 ***
## prescription_rate         -2.310e-02  2.338e-04  -98.808  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 87495  on 509  degrees of freedom
## Residual deviance: 18875  on 457  degrees of freedom
## AIC: 22843
##
## Number of Fisher Scoring iterations: 4
```



We can see that the states coefficients are all over the place - even when we notice that the expected_deaths and prescription rate vars are included. This means there is most likely something else going on in there.

Interpretations: (note that these are not perfectly valid since the variables are related with eachother so size of individual effects is approximate)

- accounting for prescription rate and expected deaths people from the state of Texas (random pick) are 55.5571941% less likely to die than the average person in the US
- accounting for state and expected deaths an increase of one in prescription rate per 100 inhabitants results in an estimated 2.2839667% lower death rate than the average person in the US
- accounting for state and prescription rate, increase in variables (national opioid mortality and state age population) leading to an increase of expected deaths by one leads to an estimated -0.0118604% decrease (so an increase by 0.0118604%) in estimated chances of dying.

States with highest mortality:

	x
stateWest Virginia	2.567154
stateKentucky	1.848173
stateTennessee	1.742284
stateOklahoma	1.496378
stateNevada	1.214991
stateOhio	1.206303
stateDelaware	1.093870
stateNew Hampshire	1.050168
stateAlabama	1.002143
stateRhode Island	0.996589

c) Population offset

by the hint - population age distribution (as well as other possible confounds) are not accounted for - old people in Florida probably die a lot more than the youth in Washington. This should be accounted for by expected deaths variable.

d) Poisson Regression - expected_deaths

```
##
## Call:
## glm(formula = deaths ~ state + prescription_rate, family = poisson,
##     data = df, offset = log(expected_deaths))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -27.0082   -2.4650    0.0442    2.4398   19.0997
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              0.1151819  0.0144904   7.949 1.88e-15 ***
## stateAlabama            -0.6031519  0.0257873 -23.390  < 2e-16 ***
## stateAlaska              0.1097643  0.0361794   3.034 0.002414 **
## stateArizona             0.0255170  0.0158007   1.615 0.106327
## stateArkansas           -0.3801609  0.0272811 -13.935  < 2e-16 ***
## stateCalifornia         -0.6630030  0.0116098 -57.107  < 2e-16 ***
## stateColorado           -0.1692181  0.0175128  -9.663  < 2e-16 ***
## stateConnecticut         0.2703020  0.0173465  15.582  < 2e-16 ***
## stateDelaware            0.3371956  0.0304757  11.064  < 2e-16 ***
## stateDistrict of Columbia  0.1567998  0.0358090   4.379 1.19e-05 ***
## stateFlorida            -0.0204810  0.0118968  -1.722 0.085149 .
## stateGeorgia            -0.3886990  0.0158693 -24.494  < 2e-16 ***
## stateHawaii             -0.7515957  0.0405336 -18.543  < 2e-16 ***
## stateIdaho              -0.5888151  0.0364016 -16.176  < 2e-16 ***
## stateIndiana            -0.2011187  0.0180568 -11.138  < 2e-16 ***
## stateIowa               -0.5500446  0.0261881 -21.004  < 2e-16 ***
## stateKansas             -0.6343163  0.0284258 -22.315  < 2e-16 ***
## stateKentucky            0.5643082  0.0179581  31.424  < 2e-16 ***
## stateLouisiana          -0.6573303  0.0245382 -26.788  < 2e-16 ***
## stateMaine               0.2902132  0.0261828  11.084  < 2e-16 ***
## stateMaryland            0.4758297  0.0136975  34.738  < 2e-16 ***
## stateMassachusetts       0.4677374  0.0132250  35.368  < 2e-16 ***
## stateMichigan            0.1311816  0.0144749   9.063  < 2e-16 ***
## stateMinnesota          -0.5309103  0.0201383 -26.363  < 2e-16 ***
## stateMississippi        -0.7952790  0.0315354 -25.219  < 2e-16 ***
## stateMissouri            0.1570852  0.0159997   9.818  < 2e-16 ***
## stateMontana            -0.4973615  0.0427850 -11.625  < 2e-16 ***
## stateNebraska           -1.2184855  0.0450237 -27.063  < 2e-16 ***
## stateNevada              0.4357719  0.0186464  23.370  < 2e-16 ***
## stateNew Hampshire       0.5847623  0.0227720  25.679  < 2e-16 ***
## stateNew Jersey         -0.1555337  0.0146629 -10.607  < 2e-16 ***
## stateNew Mexico          0.4560740  0.0203480  22.414  < 2e-16 ***
## stateNew York           -0.0861113  0.0119674  -7.195 6.22e-13 ***
## stateNorth Carolina      0.1223722  0.0142583   8.583  < 2e-16 ***
## stateNorth Dakota       -1.0126553  0.0637594 -15.882  < 2e-16 ***
## stateOhio                0.5934686  0.0125560  47.266  < 2e-16 ***
## stateOklahoma            0.3109461  0.0194118  16.018  < 2e-16 ***
## stateOregon             -0.0963428  0.0200004  -4.817 1.46e-06 ***
## statePennsylvania       -0.0355564  0.0131935  -2.695 0.007039 **
## stateRhode Island        0.5918762  0.0250275  23.649  < 2e-16 ***
## stateSouth Carolina     -0.1364762  0.0194754  -7.008 2.42e-12 ***
## stateSouth Dakota       -0.8681808  0.0560265 -15.496  < 2e-16 ***
## stateTennessee           0.3000179  0.0183596  16.341  < 2e-16 ***
## stateTexas              -0.7683252  0.0128625 -59.734  < 2e-16 ***
## stateUtah                0.4476126  0.0182937  24.468  < 2e-16 ***
## stateVermont             0.1168908  0.0394756   2.961 0.003066 **
## stateVirginia           -0.1361692  0.0150069  -9.074  < 2e-16 ***
## stateWashington          0.0028164  0.0152289   0.185 0.853279
## stateWest Virginia       1.1421947  0.0199774  57.174  < 2e-16 ***
## stateWisconsin           0.0565688  0.0159898   3.538 0.000403 ***
## stateWyoming            -0.1844081  0.0479815  -3.843 0.000121 ***
## prescription_rate       -0.0006780  0.0001908  -3.554 0.000379 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 63933  on 509  degrees of freedom
## Residual deviance: 13675  on 458  degrees of freedom
## AIC: 17641
##
## Number of Fisher Scoring iterations: 4
```

Previously the interpretation of $(1 - exp(\text{coefficient})) * 100\%$ was that the variable was associated with that % decrease in mortality compared to the average person in the population assuming that all populations are the same in distribution just not in number.

Now this is gonna take into account that the populations have different distributions.

e) Overdispersion

Let's look at mean and sd of residuals, these should be 0 and 1 respectively for not overdispersed data.

mean	sd
0.083236	5.093708

These definitely don't have standard deviation of 1!

Estimated overdispersion factor:

The overdispersion factor is 27.5231973 which means that the standard errors are inflated by 5.2462556 which is quite a bit. So yeah, there is an issue.

f) Negative Binomial Regression

```
##
## Call:
## MASS::glm.nb(formula = deaths ~ expected_deaths + state + prescription_rate,
##     data = df, init.theta = 14.6212095, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.3532  -0.6930  -0.0417   0.5206   3.9614
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                7.966e+00  1.593e-01  50.015  < 2e-16 ***
## expected_deaths            2.986e-04  4.883e-05   6.115 9.63e-10 ***
## stateAlabama              -2.730e-02  1.542e-01  -0.177 0.859491
## stateAlaska               -2.308e+00  1.327e-01 -17.395  < 2e-16 ***
## stateArizona              -1.027e-01  1.214e-01  -0.846 0.397630
## stateArkansas             -4.833e-01  1.422e-01  -3.399 0.000676 ***
## stateCalifornia           -3.180e-01  1.595e-01  -1.994 0.046180 *
## stateColorado             -6.667e-01  1.217e-01  -5.479 4.27e-08 ***
## stateConnecticut          -7.225e-01  1.243e-01  -5.812 6.19e-09 ***
## stateDelaware             -1.412e+00  1.306e-01 -10.817  < 2e-16 ***
## stateDistrict of Columbia -2.856e+00  1.468e-01 -19.460  < 2e-16 ***
## stateFlorida               5.231e-01  1.262e-01   4.146 3.38e-05 ***
## stateGeorgia              -9.548e-02  1.225e-01  -0.779 0.435710
## stateHawaii               -2.897e+00  1.399e-01 -20.700  < 2e-16 ***
## stateIdaho                -1.912e+00  1.303e-01 -14.677  < 2e-16 ***
## stateIndiana              -3.764e-02  1.290e-01  -0.292 0.770537
## stateIowa                 -1.615e+00  1.264e-01 -12.772  < 2e-16 ***
## stateKansas               -1.408e+00  1.268e-01 -11.107  < 2e-16 ***
## stateKentucky              8.376e-01  1.421e-01   5.892 3.81e-09 ***
## stateLouisiana            -5.733e-01  1.354e-01  -4.236 2.28e-05 ***
## stateMaine                -1.311e+00  1.279e-01 -10.254  < 2e-16 ***
## stateMaryland             -2.791e-02  1.208e-01  -0.231 0.817255
## stateMassachusetts        -5.816e-02  1.208e-01  -0.481 0.630271
## stateMichigan              5.348e-01  1.261e-01   4.240 2.23e-05 ***
## stateMinnesota            -1.283e+00  1.248e-01 -10.284  < 2e-16 ***
## stateMississippi          -9.803e-01  1.407e-01  -6.969 3.19e-12 ***
## stateMissouri              1.121e-01  1.238e-01   0.906 0.365110
## stateMontana              -2.297e+00  1.321e-01 -17.380  < 2e-16 ***
## stateNebraska             -2.723e+00  1.329e-01 -20.488  < 2e-16 ***
## stateNevada               -1.529e-01  1.267e-01  -1.207 0.227515
## stateNew Hampshire        -1.087e+00  1.269e-01  -8.564  < 2e-16 ***
## stateNew Jersey           -5.405e-01  1.198e-01  -4.512 6.43e-06 ***
## stateNew Mexico           -8.613e-01  1.257e-01  -6.850 7.36e-12 ***
## stateNew York             -1.303e-01  1.208e-01  -1.079 0.280551    
## stateNorth Carolina        4.949e-01  1.247e-01   3.968 7.23e-05 ***
## stateNorth Dakota         -3.649e+00  1.455e-01 -25.083  < 2e-16 ***
## stateOhio                  1.001e+00  1.257e-01   7.966 1.63e-15 ***
## stateOklahoma              4.100e-01  1.375e-01   2.981 0.002871 ** 
## stateOregon               -3.633e-01  1.264e-01  -2.874 0.004048 ** 
## statePennsylvania          2.078e-01  1.203e-01   1.728 0.084017 .  
## stateRhode Island         -1.328e+00  1.278e-01 -10.386  < 2e-16 ***
## stateSouth Carolina       -2.444e-01  1.280e-01  -1.909 0.056199 .  
## stateSouth Dakota         -3.376e+00  1.419e-01 -23.788  < 2e-16 ***
## stateTennessee             1.042e+00  1.495e-01   6.974 3.07e-12 ***
## stateTexas                -2.197e-01  1.347e-01  -1.632 0.102762    
## stateUtah                 -3.997e-01  1.246e-01  -3.208 0.001334 ** 
## stateVermont              -2.668e+00  1.371e-01 -19.461  < 2e-16 ***
## stateVirginia             -2.446e-01  1.193e-01  -2.049 0.040433 *  
## stateWashington           -7.117e-02  1.204e-01  -0.591 0.554471    
## stateWest Virginia         7.517e-01  1.479e-01   5.083 3.72e-07 ***
## stateWisconsin            -3.852e-01  1.212e-01  -3.180 0.001475 ** 
## stateWyoming              -2.608e+00  1.346e-01 -19.380  < 2e-16 ***
## prescription_rate         -2.016e-02  1.606e-03 -12.548  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(14.6212) family taken to be 1)
## 
##     Null deviance: 8277.04  on 509  degrees of freedom
## Residual deviance:  521.15  on 457  degrees of freedom
## AIC: 6097.2
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  14.621 
##           Std. Err.:  0.984 
## 
##  2 x log-likelihood:  -5989.177
```

It does change quite a few of significaces down to not-significant. The states could probably be grouped more into buckets of states either by region or by population distribution. As expected.

LRT:

P-value is 0 (pretty much zero) so NB is much better. This is also very clear from looking at raw likelihood numbers - the difference is in the thousands while the df difference is exactly 1 (the theta for NB)

g) Summary

After iterating through a bunch of models (none of which were a very good fit) it's clear that the expected number of deaths is predictive of mortality, the mortality is also highly variable per state, due to things other than prescription rate and population size and distribution as well as unemployment. Presecription rate seems to have an effect. Overall there definitely is a need for both: a more advanced modelling approach and a larger number of variables that could help account for the inter-state variation.