Assignment 1 - STA2201H Applied Statistics II

Luis Correia - Student No. 1006508566

January 17th 2020

Question 1 - Exponential family

The random variable Y belongs to the exponential family of distributions if its support does ot depend upon any unknown parameters and its density or probability mass function takes the form:

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

Assume ϕ is known.

(a) Show $\int \left[\frac{dp}{d\theta}\right] = 0$ and $\int \left[\frac{d^2p}{d\theta^2}\right] = 0$.

{*Solution.*}

We know that, as per definition, since $p(y|\theta,\phi)$ is a density we have:

$$\int p(y|\theta,\phi)dy = 1$$

Deriving both sides in relation to θ we have:

$$\frac{d}{d\theta}\int p(y|\theta,\phi)dy = \frac{d}{d\theta}1$$

$$\implies \int \frac{d}{d\theta}p(y|\theta,\phi)dy = 0. \tag{1}$$

The part 2 of this problem, we have:

$$\int \frac{d^2}{d\theta^2}p(y|\theta,\phi)dy = \int \frac{d}{d\theta}\left(\frac{d}{d\theta}p(y|\theta,\phi)dy\right) = \frac{d}{d\theta}\int \frac{d}{d\theta}p(y|\theta,\phi)dy$$

Using the result from (1), we have:

$$\frac{d}{d\theta}0 = 0 \implies \int \frac{d^2}{d\theta^2}p(y|\theta,\phi)dy = 0. \tag{2}$$

(b) Using a) Show $E[Y] = b'(\theta)$ and $Var[Y] = \phi b''(\theta)$.

{*Solution.*}

Considering that Y is from Exponential family, we have:

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

Deriving both sides in relation to θ, we have:

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

$$\implies \frac{d}{d\theta}p(y|\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)\Big\{\phi^{-1}\left[y - b'(\theta)\right]\Big\}$$

$$\implies \frac{d}{d\theta}p(y|\theta,\phi) = p(y|\theta,\phi)\Big\{\phi^{-1}\left[y - b'(\theta)\right]\Big\}$$

Integrating both sides in relation to $\boldsymbol{y}$ we have:

$$\implies \int \frac{d}{d\theta}p(y|\theta,\phi)dy = \int \Big[p(y|\theta,\phi)\Big\{\phi^{-1}\left[y - b'(\theta)\right]\Big\}\Big]dy \tag{3}$$

From (1), the left side of (3) is equal to zero, then

$$\begin{aligned}
\implies 0 &= \int \Big[p(y|\theta,\phi)\Big\{\phi^{-1}\left[y - b'(\theta)\right]\Big\}\Big]dy \\
&= \phi^{-1}\Big[\int \big(y - b'(\theta)\big)p(y|\theta,\phi)dy\Big] \\
&= \phi^{-1}\Big[\int y.p(y|\theta,\phi)dy - b'(\theta)\int p(y|\theta,\phi)dy\Big] \\
&= \phi^{-1}\big[E(Y) - b'(\theta).1\big] = 0
\end{aligned}$$

$$\implies E(Y) - b'(\theta) = 0$$

$$\implies E(Y) = b'(\theta). \tag{4}$$

The part 2 of this problem, we have:

$$\begin{aligned}
\frac{d^2}{d\theta^2}p(y|\theta,\phi) &= \frac{d}{d\theta}\Big[\frac{d}{d\theta}p(y|\theta,\phi)\Big] \\
&= \frac{d}{d\theta}\Big[p(y|\theta,\phi)\big\{\phi^{-1}\left[y - b'(\theta)\right]\big\}\Big] \\
&= \phi^{-1}\frac{d}{d\theta}\Big[p(y|\theta,\phi)\left[y - b'(\theta)\right]\Big] \\
&= \phi^{-1}\frac{d}{d\theta}\Big[y.p(y|\theta,\phi) - b'(\theta).p(y|\theta,\phi)\Big] \\
&= \phi^{-1}\Big[y\frac{d}{d\theta}p(y|\theta,\phi)\Big] - \phi^{-1}\Big[\frac{d}{d\theta}b'(\theta).p(y|\theta,\phi)\Big]
\end{aligned}$$

Integrating both sides in relation to $\boldsymbol{y}$ we have:

$$\int \frac{d^2}{d\theta^2} p(y|\theta,\phi)dy = \int \phi^{-1}\Big[y\frac{d}{d\theta}p(y|\theta,\phi)\Big]dy - \int \phi^{-1}\Big[\frac{d}{d\theta}b'(\theta).p(y|\theta,\phi)\Big]dy$$

From (2) we know the left side of the equation is equal zero, then we have:

$$\begin{aligned}
0 &= \phi^{-1}\Big\{\int \Big[y\frac{d}{d\theta}p(y|\theta,\phi)\Big]dy - \int \Big[\frac{d}{d\theta}b'(\theta).p(y|\theta,\phi)\Big]dy\Big\} \\
&= \phi^{-1}\Big\{\int y.p(y|\theta,\phi)\phi^{-1}\left[y-b'(\theta)\right]dy - \int \Big[\frac{d}{d\theta}b'(\theta).p(y|\theta,\phi)\Big]dy\Big\} \\
&= \phi^{-1}\Big\{\phi^{-1}\int y^2.p(y|\theta,\phi)dy - \phi^{-1}b'(\theta)\int y.p(y|\theta,\phi)dy - \int \Big[\frac{d}{d\theta}b'(\theta).p(y|\theta,\phi)\Big]dy\Big\} \\
&= \phi^{-1}\Big\{\phi^{-1}E(Y^2) - \phi^{-1}b'(\theta)E(Y) - \int \Big[b''(\theta).p(y|\theta,\phi) + b'(\theta).p(y|\theta,\phi)\phi^{-1}\left[y-b'(\theta)\right]\Big]dy\Big\} \\
&= \phi^{-1}\Big\{\phi^{-1}E(Y^2) - \phi^{-1}b'(\theta)E(Y) - b''(\theta)\int p(y|\theta,\phi)dy + b'(\theta)\int p(y|\theta,\phi)\phi^{-1}\left[y-b'(\theta)\right]dy\Big\} \\
&= \phi^{-1}\Big\{\phi^{-1}E(Y^2) - \phi^{-1}b'(\theta)E(Y) - b''(\theta).1 + \phi^{-1}b'(\theta)\int y.p(y|\theta,\phi)dy - \phi^{-1}b'(\theta)^2\int p(y|\theta,\phi)dy\Big\} \\
&= \phi^{-1}\Big\{\phi^{-1}E(Y^2) - \phi^{-1}b'(\theta)E(Y) - b''(\theta) + \phi^{-1}b'(\theta)E(Y) - \phi^{-1}b'(\theta)^2.1\Big\} \\
&= \phi^{-1}\Big\{\phi^{-1}E(Y^2) - \phi^{-1}\left[E(Y)\right]^2 - b''(\theta)\Big\} \\
&= \phi^{-1}\Big\{\phi^{-1}Var(Y) - b''(\theta)\Big\} = 0
\end{aligned}$$

$$\implies \phi^{-1}Var(Y) - b''(\theta) = 0$$

$$\implies Var(Y) = \phi b''(\theta). \tag{5}$$

(c) Denote $\log p(y|\theta,\phi)$ as $\ell(\theta)$. Using (b) show $\mathrm{E}\big[\frac{\partial \ell(\theta)}{\partial\theta}\big] = 0$ and $\mathrm{Var}\big[\frac{\partial \ell(\theta)}{\partial\theta}\big] = \phi^{-1}b''(\theta)$.

{*Solution.*}

$$\begin{aligned}
\frac{\partial \ell(\theta)}{\partial\theta} &= \frac{\partial}{\partial\theta}\left\{log\left[exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)\right]\right\} \\
&= \frac{\partial}{\partial\theta}\left[\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right] \\
&= \phi^{-1}\left[y - b'(\theta)\right]
\end{aligned}$$

Then calculating the $\mathrm{E}\big[\frac{\partial \ell(\theta)}{\partial\theta}\big]$ and using (4) we have:

$$\begin{aligned}
E\Big[\frac{\partial \ell(\theta)}{\partial\theta}\Big] &= E\big\{\phi^{-1}\left[Y - b'(\theta)\right]\big\} \\
&= \phi^{-1}\big\{E\left[Y - b'(\theta)\right]\big\} \\
&= \phi^{-1}\left[E(Y) - b'(\theta)\right] = 0
\end{aligned}$$

$$\implies E\Big[\frac{\partial \ell(\theta)}{\partial\theta}\Big] = 0. \tag{6}$$

Calculating $\mathrm{Var}\big[\frac{\partial \ell(\theta)}{\partial\theta}\big]$ we have:

$$\begin{aligned} Var\Big[\frac{\partial \ell(\theta)}{\partial \theta}\Big] &= Var\big\{\phi^{-1}\left[Y - b'(\theta)\right]\big\} \\ &= \phi^{-2} Var\big[Y - b'(\theta)\big] \\ &= \phi^{-2} Var(Y) \end{aligned}$$

Using the result (5) we have then:

$$\implies Var\Big[\frac{\partial \ell(\theta)}{\partial \theta}\Big] = \phi^{-1} b''(\theta). \tag{7}$$

Question 2 - Overdispersion

Suppose that the conditional distribution of outcome Y given an unobserved variable θ is Poisson, with a mean and variance $\mu\theta$, so:

$$Y|\theta \sim Poisson(\mu\theta) \tag{8}$$

(a) Assume E(θ) = 1 and Var(θ) = σ^2. Using the laws of total expectation and total variance, show E(Y) = μ and Var(Y) = $\mu(1 + \mu\sigma^2)$.

{*Solution.*}

Law of Total Expectation

$$E(Y) = E(E(Y|X)) \tag{9}$$

and

Law of Total Variance

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X)) \tag{10}$$

Using (9) and (8) we have:

$$\begin{aligned} E(Y) = E(E(Y|\theta)) = E(\mu\theta) = \\ = \mu E(\theta) = \\ = \mu. \end{aligned}$$

Using (10) and (8) we have:

$$\begin{aligned} Var(Y) &= E(Var(Y|\theta)) + Var(E(Y|\theta)) = \\ &= E(\mu\theta) + Var(\mu\theta) = \\ &= \mu + \mu^2 Var(\theta) = \\ &= \mu + \mu^2\sigma^2 = \\ &= \mu\left(1 + \mu\sigma^2\right). \end{aligned}$$

(b) Assume E(θ) is Gamma distributed with α and β as shape and scale parameters, respectively. Show the unconditional distribution of Y is Negative Binomial.

{*Solution.*}

Let $Y|\theta \sim$ Poisson($\mu\theta$) and $\Theta \sim$ Gamma(α,β).

The joint p.d.f. of Y and Θ is given by:

$$\begin{aligned} f_{Y,\Theta}(y,\theta) &= P(Y = y|\Theta = \theta)P(\Theta = \theta) \\ &= e^{-\mu\theta}\frac{(\mu\theta)^y}{y!}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \end{aligned}$$

The marginal distribution (or unconditional distribution) of Y is calculated by integrating the joint p.d.f. $f_{Y,\Theta}(y,\theta)$ over all values of Θ, then we have:

$$\begin{aligned} P(Y=y) &= \int_0^\infty f_{Y,\Theta}(y,\theta)d\theta \\ &= \int_0^\infty e^{-\mu\theta}\frac{(\mu\theta)^y}{y!}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\mu^y}{y!}\int_0^\infty e^{-\mu\theta}\theta^y\theta^{\alpha-1}e^{-\beta\theta}d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\mu^y}{y!}\int_0^\infty \theta^{y+\alpha-1}e^{-(\mu+\beta)\theta}d\theta \end{aligned}$$

Multiplying this equation by $\frac{\Gamma(y+\alpha)}{(\beta+\mu)^{y+\alpha}}\frac{(\beta+\mu)^{y+\alpha}}{\Gamma(y+\alpha)}$ we have the following:

$$\implies P(Y=y) = \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\mu^y}{y!}\frac{\Gamma(y+\alpha)}{(\beta+\mu)^{y+\alpha}}\int_0^\infty \frac{(\beta+\mu)^{y+\alpha}}{\Gamma(y+\alpha)}\theta^{y+\alpha-1}e^{-(\mu+\beta)\theta}d\theta \tag{11}$$

Please note that the function at right of the integral sign in (11) is the p.d.f. of a $Gamma(y+\alpha, \beta+\mu)$, then this expression is equal to 1.

Then (11) reduces to:

$$\begin{aligned} P(Y=y) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\mu^y}{y!}\frac{\Gamma(y+\alpha)}{(\beta+\mu)^{y+\alpha}} \\ &= \frac{\Gamma(y+\alpha)}{\Gamma(y+1)\Gamma(\alpha)}\left(\frac{\beta}{\beta+\mu}\right)^\alpha\left(\frac{\mu}{\beta+\mu}\right)^y \\ &= \frac{(y+\alpha-1)!}{(\alpha-1)!y!}\left(\frac{\beta}{\beta+\mu}\right)^\alpha\left(1-\frac{\beta}{\beta+\mu}\right)^y \end{aligned}$$

$$\implies P(Y=y) = \binom{\alpha+y-1}{y}\left(\frac{\beta}{\beta+\mu}\right)^\alpha\left(1-\frac{\beta}{\beta+\mu}\right)^y \tag{12}$$

Then from (12) we conclude that Y has a p.d.f Negative Binomial with parameters $r=\alpha$ and $p=\frac{\beta}{\beta+\mu}$.

(c) In order for E(Y) = μ and Var(Y) = $\mu(1+\mu\sigma^2)$, what must α and β equal?

{*Solution.*}

In part (b) above we concluded that $Y \sim NB(r,p)$ then we have:

$$\implies E(Y) = \frac{r(1-p)}{p} \tag{13}$$

and

$$\implies Var(Y) = \frac{r(1-p)}{p^2} \tag{14}$$

We also know from (12) that the parameters of Y are $r=\alpha$ and $p=\frac{\beta}{\beta+\mu}$, then replacing this values in equations (13) and (14) we have:

$$\implies E(Y) = \frac{r(1-p)}{p} = \frac{\alpha\left(1-\frac{\beta}{\beta+\mu}\right)}{\frac{\beta}{\beta+\mu}} = \frac{\alpha\mu}{\beta} \tag{15}$$

and

$$\implies Var(Y) = \frac{r(1-p)}{p^2} = \frac{\alpha\left(1-\frac{\beta}{\beta+\mu}\right)}{\left(\frac{\beta}{\beta+\mu}\right)^2} = \frac{\alpha\mu\left(\beta+\mu\right)}{\beta^2} \tag{16}$$

In our case, $E(Y) = \mu$ and $Var(Y) = \mu\left(1+\mu\sigma^2\right)$, so by replacing these values in equations (15) and (16) we obtain the desired values for α and β as follows:

$$\implies \frac{\alpha\mu}{\beta} = \mu \implies \alpha = \beta \tag{17}$$

and

$$\implies \frac{\alpha\mu\left(\beta+\mu\right)}{\beta^2} = \mu\left(1+\mu\sigma^2\right) \tag{18}$$

Replacing (17) in (18) we have

$$\begin{aligned}
\frac{\beta\mu\left(\beta+\mu\right)}{\beta^2} &= \mu\left(1+\mu\sigma^2\right) \\
\frac{\beta+\mu}{\beta} &= 1+\mu\sigma^2 \\
\beta+\mu &= \beta+\beta\mu\sigma^2 \\
\mu &= \beta\mu\sigma^2 \\
1 &= \beta\sigma^2 \\
\implies \beta &= \frac{1}{\sigma^2} = \alpha.
\end{aligned} \tag{19}$$

Question 3 - Simulation

Generate 100 datasets of an explanatory x and Poisson outcome y with the following code:

```
set.seed(123)
N <- 100                # No. of Samples
Size <- 100             # Sample Size
X <- matrix(NA, N, Size)
Y <- matrix(NA, N, Size)

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

- (a) Fit a Poisson GLM to each dataset, with the 'correct' explanatory variables (x and x^2). Store the coefficient estimates and standard errors from each model run (hint: you can get SEs from **sqrt(diag(vcov(mod)))** where mod is your model object).

```
realparam <- as.numeric(c(0.5,   # Intercept
                          1.0,   # Coef. X (=beta0)
                          0.2))  # Coef. X^2 (=beta1)

P <- length(realparam)  # No. Of parameters in glm() model

XQd <- matrix(NA, N, Size)
CFs <- matrix(NA, N, P)
SEs <- matrix(NA, N, P)
VHs <- array(data = NA, dim = c(N, dim(matrix(NA, P, P))))

# Stores best fit on 1st simmulation
bestfit1 <- glm(formula = Y[1,]~X[1,]+(XQd[1,] <- X[1,]^2), family=poisson())
worstfit1 <- bestfit1   # Start value

cf <- coef(bestfit1)
se <- sqrt(diag(vcov(bestfit1)))
vh <- vcov(bestfit1)
CFs[1,] <- cf
SEs[1,] <- se
VHs[1,,] <- vh


for(i in 2:N){
  fit.correct <- glm(formula = Y[i,]~X[i,]+(XQd[i,] <- X[i,]^2), family=poisson())
  cf <- coef(fit.correct)
  se <- sqrt(diag(vcov(fit.correct)))
  vh <- vcov(fit.correct)
  CFs[i,] <- cf
  SEs[i,] <- se
  VHs[i,,] <- vh
  if (AIC(fit.correct)< AIC(bestfit1))
    bestfit1 <- fit.correct
  if (AIC(fit.correct)> AIC(worstfit1))
```

```
    worstfit1 <- fit.correct
}
```

- (b) Calculate the coverage probability of a 2 standard error confidence interval for the coefficient on x and assess whether this is a useful way to construct 95

```
conf_interval_2SE <- function(p) { # p[1] = Coeficient; p[2] = SE
  upper <- p[1] + 2.0*p[2] #find the upper bound for a 2*SE CI
  lower <- p[1] - 2.0*p[2] #find the lower bound for a 2*SE CI
  return(c(lower,upper))
}

#check if the interval contains the true coefficient

interval_contains_true_coef <- function(p) {
  p[3] >= p[1] && p[3] <= p[2]
}

#Finds the confidence intervals for beta0 (2*SE CI)

intervals2SE <- t(apply(cbind(CFs[,2], SEs[,2]), FUN=conf_interval_2SE, MARGIN=1))

colnames(intervals2SE) <- c("lower","upper")

dt2SE <- cbind.data.frame(seq(1:N),
                          intervals2SE[,1],
                          intervals2SE[,2],
                          CFs[,2])
colnames(dt2SE) <- c("id", "lower", "upper", "coefhat"); dt2SE
```

```
##       id     lower     upper   coefhat
## 1      1 0.9287548 1.4249428 1.1768488
## 2      2 0.8890034 1.2260971 1.0575503
## 3      3 0.9106510 1.2885068 1.0995789
## 4      4 0.7349787 1.0586085 0.8967936
## 5      5 0.7748969 1.1096275 0.9422622
## 6      6 0.7155708 1.0159167 0.8657438
## 7      7 0.8630456 1.2114751 1.0372603
## 8      8 0.7737082 1.0839030 0.9288056
## 9      9 0.7388984 1.0697138 0.9043061
## 10    10 0.7710357 1.1158834 0.9434596
## 11    11 0.8737916 1.1973820 1.0355868
## 12    12 0.7779745 1.1701451 0.9740598
## 13    13 0.8447810 1.1037609 0.9742709
## 14    14 0.7943448 1.0959659 0.9451554
## 15    15 0.8112388 1.1457161 0.9784775
## 16    16 0.8047262 1.1326421 0.9686842
## 17    17 0.8099614 1.2043067 1.0071340
## 18    18 0.8212332 1.1285987 0.9749159
## 19    19 0.6744357 0.9948053 0.8346205
## 20    20 0.7754743 1.1330675 0.9542709
## 21    21 0.8662107 1.2310895 1.0486501
## 22    22 0.9014275 1.3296743 1.1155509
## 23    23 0.9339383 1.3624030 1.1481707
## 24    24 0.8282770 1.2051348 1.0167059
```

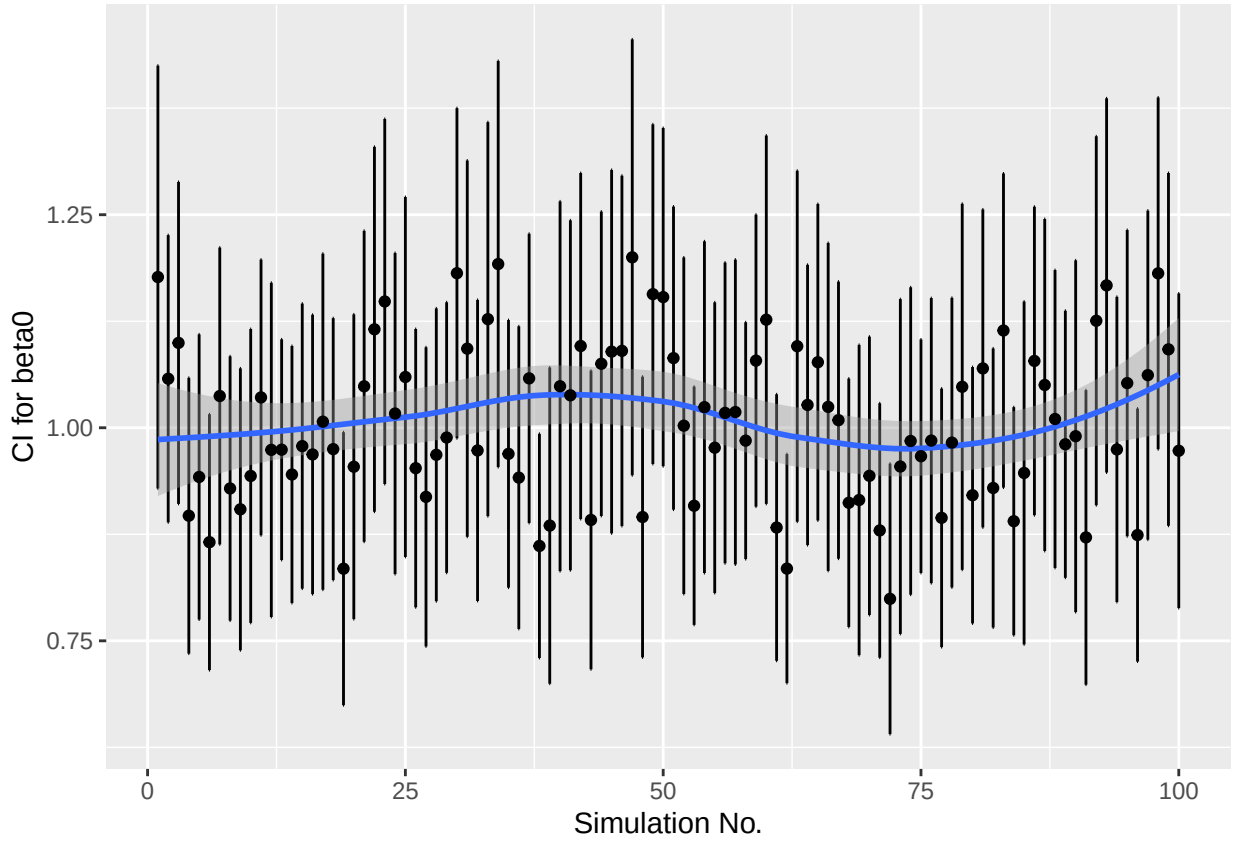
```
## 25   25 0.8483709 1.2707978 1.0595844
## 26   26 0.7892979 1.1159052 0.9526015
## 27   27 0.7433498 1.0945089 0.9189293
## 28   28 0.7964458 1.1400843 0.9682651
## 29   29 0.8297956 1.1472532 0.9885244
## 30   30 0.9876401 1.3750901 1.1813651
## 31   31 0.8720177 1.3136212 1.0928195
## 32   32 0.7967219 1.1500225 0.9733722
## 33   33 0.8962259 1.3586351 1.1274305
## 34   34 0.9538397 1.4304039 1.1921218
## 35   35 0.8123940 1.1264128 0.9694034
## 36   36 0.7639229 1.1190628 0.9414928
## 37   37 0.8882530 1.2276834 1.0579682
## 38   38 0.7297875 0.9928924 0.8613400
## 39   39 0.6996438 1.0707981 0.8852209
## 40   40 0.8317398 1.2656874 1.0487136
## 41   41 0.8327501 1.2434110 1.0380806
## 42   42 0.8928760 1.2989757 1.0959258
## 43   43 0.7165050 1.0671093 0.8918071
## 44   44 0.8964646 1.2537288 1.0750967
## 45   45 0.8760657 1.3025664 1.0893160
## 46   46 0.8846619 1.2958147 1.0902383
## 47   47 0.9442372 1.4557678 1.2000025
## 48   48 0.7305405 1.0599683 0.8952544
## 49   49 0.9574678 1.3560546 1.1567612
## 50   50 0.9548473 1.3518190 1.1533332
## 51   51 0.9039272 1.2595966 1.0817619
## 52   52 0.8049379 1.1999095 1.0024237
## 53   53 0.7686883 1.0482764 0.9084824
## 54   54 0.8295153 1.2188367 1.0241760
## 55   55 0.8060375 1.1471592 0.9765984
## 56   56 0.8409579 1.1939088 1.0174333
## 57   57 0.8395482 1.1973527 1.0184504
## 58   58 0.8459076 1.1238333 0.9848704
## 59   59 0.9074221 1.2501456 1.0787838
## 60   60 0.9106921 1.3429119 1.1268020
## 61   61 0.7265727 1.0393143 0.8829435
## 62   62 0.7002936 0.9691880 0.8347408
## 63   63 0.8897572 1.3016286 1.0956929
## 64   64 0.8622824 1.1912504 1.0267664
## 65   65 0.8912008 1.2626238 1.0769123
## 66   66 0.8320862 1.2168726 1.0244794
## 67   67 0.8463510 1.1715262 1.0089386
## 68   68 0.7662376 1.0577431 0.9119904
## 69   69 0.7331483 1.0973059 0.9152271
## 70   70 0.7802470 1.1069656 0.9436063
## 71   71 0.7302594 1.0287771 0.8795182
## 72   72 0.6404722 0.9579602 0.7992162
## 73   73 0.7580824 1.1509474 0.9545149
## 74   74 0.8041341 1.1648392 0.9844867
## 75   75 0.8295536 1.1036835 0.9666186
## 76   76 0.8176466 1.1522100 0.9849283
## 77   77 0.7426183 1.0460191 0.8943187
## 78   78 0.8126020 1.1524981 0.9825500
```

```
## 79   79 0.8333122 1.2628567 1.0480844
## 80   80 0.7703821 1.0712991 0.9208406
## 81   81 0.8830545 1.2562616 1.0696580
## 82   82 0.7655307 1.0931205 0.9293256
## 83   83 0.9297715 1.2984494 1.1141105
## 84   84 0.7565074 1.0242040 0.8903557
## 85   85 0.7457588 1.1481355 0.9469472
## 86   86 0.8973606 1.2594353 1.0783979
## 87   87 0.8554216 1.2449545 1.0501881
## 88   88 0.8355118 1.1850757 1.0102937
## 89   89 0.8239260 1.1373507 0.9806384
## 90   90 0.7836975 1.1964417 0.9900696
## 91   91 0.6985340 1.0441725 0.8713533
## 92   92 0.9090786 1.3420522 1.1255654
## 93   93 0.9474672 1.3866175 1.1670423
## 94   94 0.7955114 1.1536308 0.9745711
## 95   95 0.8726445 1.2320938 1.0523692
## 96   96 0.7256569 1.0225833 0.8741201
## 97   97 0.8686836 1.2548100 1.0617468
## 98   98 0.9748014 1.3875661 1.1811838
## 99   99 0.8849794 1.2989832 1.0919813
## 100 100 0.7883880 1.1576624 0.9730252
```

```
# Plot

ggplot(dt2SE, aes(x=dt2SE$id, y=dt2SE$coefhat, group=1)) +
  geom_errorbar(aes(ymin=dt2SE$lower, ymax=dt2SE$upper, width=.1)) +
  geom_smooth() +
  geom_point() +
  xlab("Simulation No.") + ylab("CI for beta0")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
percent_intervals_with_true_coef <- apply(cbind(dt2SE[,2:3],rep(realparam[2],N)), FUN=interval_contains_

cat("% Coverage Probability for coefficient of 'x' : ", sum(percent_intervals_with_true_coef)/N*100, "%"
```

```
## % Coverage Probability for coefficient of 'x' :  96 %
```

Apparently this technique is useful to obtain 95% confidence intervals since the coverage probability is very close to the desired level. Suggestion: increase the quantity of samples (N) and size of each sample (Size) and verify of the behaviour still holds.

- (c) Calculate 100 Wald tests for the coefficient on x against $\beta_0 = 1$ (i.e. the true value). (hint: the pt() function will return the P-value of a t-test with a specified degrees of freedom). In how many of the tests is the null rejected?

```
### This function was developed by Prof. Jerry Brunner and applied on Wald-Tests in Applied Stats I ###
### Its usage is permitted by Prof. Brunner ###

Wtest  <-  function(L,Tn,Vn,h=0) # H0: L theta = h
  # Tn is estimated theta, usually a vector.
  # Vn is the estimated asymptotic covariance matrix of Tn.
  # For Wald tests based on numerical MLEs, Tn = theta-hat,
  # and Vn is the inverse of the Hessian of the minus log
  # likelihood.
{
  Wtest  <- numeric(4)
  names(Wtest) <- c("W","df","p-value", "h")
  r  <-  dim(L)[1]
  W <- t(L%*%Tn-h) %*% solve(L%*%Vn%*%t(L)) %*%
```

```
    (L%*%Tn-h)
  W<- as.numeric(W)
  pval <- 1-pchisq(W,r)
  Wtest[1] <- W; Wtest[2] <- r; Wtest[3]  <-  pval; Wtest[4] <- h
  Wtest
} # End function Wtest

# For Wald tests: Wtest = function(L,Tn,Vn,h=0) # H0: L theta = h

WT <- array(data = NA, dim = c(N,4))
dimnames(WT)[[2]] <- list("W", "df", "p-value", "h")

# Testing H0: beta0 = 1 (h)

L0 <- rbind(c(0,1,0))

cat("\n Wald-Test\n")
```

```
##
##  Wald-Test
```

```
for(i in 1:N)
  WT[i,] <- Wtest(L0, CFs[i,], VHs[i,,], realparam[2])

WTReject <- which(WT[,"p-value"]<0.05)

cat("\n\n No. of Rejections of Ho: ", length(WTReject))
```

```
##
##
##  No. of Rejections of Ho:  4
```

Now generate 100 new datasets based on the code below:

```
set.seed(321)
X2 <- matrix(NA, N, Size)
Y2 <- matrix(NA, N, Size)

for(i in 1:N){
  weights <- ifelse(X[i,]>1, 10, 1)
  probs <- weights/sum(weights)
  to_keep_2 <- sample(1:length(X[i,]), 25, prob = probs)
  x2 <- X[i,to_keep_2]
  y2 <- Y[i,to_keep_2]
  X2[i,] <- x2
  Y2[i,] <- y2
}
```

- (d) Repeat parts a)-c) on the new data X2 and Y2.

REPEAT Part (a): Fitting **glm()** model

```
bestfit2 <- glm(formula = Y2[1,]~X2[1,]+(XQd[1,] <- X2[1,]^2), family=poisson())  # Stores best fit on
worstfit2 <- bestfit2

cf <- coef(bestfit2)
se <- sqrt(diag(vcov(bestfit2)))
```

```
vh <- vcov(bestfit2)
CFs[1,] <- cf
SEs[1,] <- se
VHs[1,,] <- vh

for(i in 2:N){
  fit2.correct <- glm(formula = Y2[i,]~X2[i,]+(XQd[i,] <- X2[i,]^2), family=poisson())
  cf <- coef(fit2.correct)
  se <- sqrt(diag(vcov(fit2.correct)))
  vh <- vcov(fit2.correct)
  CFs[i,] <- cf
  SEs[i,] <- se
  VHs[i,,] <- vh
  if (AIC(fit2.correct)< AIC(bestfit2))
    bestfit2 <- fit2.correct
  if (AIC(fit2.correct)> AIC(worstfit2))
    worstfit2 <- fit2.correct
}
```

REPEAT Part (b): Coverage Probability of Confidence Interval

```
intervals2SE <- t(apply(cbind(CFs[,2], SEs[,2]), FUN=conf_interval_2SE, MARGIN=1)) #Finds the confidenc

colnames(intervals2SE) <- c("lower","upper")

dt2SE <- cbind.data.frame(seq(1:N),
                          intervals2SE[,1],
                          intervals2SE[,2],
                          CFs[,2])
colnames(dt2SE) <- c("id", "lower", "upper", "coefhat"); dt2SE
```

```
##       id     lower     upper   coefhat
## 1      1 0.8238894 1.3681432 1.0960163
## 2      2 0.9729732 1.3879178 1.1804455
## 3      3 0.7916221 1.2687247 1.0301734
## 4      4 0.5077987 0.8625549 0.6851768
## 5      5 0.8623852 1.4134212 1.1379032
## 6      6 0.6412702 1.0254282 0.8333492
## 7      7 0.8080398 1.1977988 1.0029193
## 8      8 0.7005702 1.1138857 0.9072279
## 9      9 0.7721680 1.2877398 1.0299539
## 10    10 0.7030308 1.0953815 0.8992062
## 11    11 0.9118780 1.2347572 1.0733176
## 12    12 0.5037165 0.8569974 0.6803569
## 13    13 0.7017812 0.9788385 0.8403099
## 14    14 0.5672167 0.9765124 0.7718645
## 15    15 0.9780088 1.4329852 1.2054970
## 16    16 0.9256690 1.2753941 1.1005316
## 17    17 0.6457362 1.0508258 0.8482810
## 18    18 0.9556538 1.4415197 1.1985867
## 19    19 0.8177832 1.2484641 1.0331237
## 20    20 0.9862170 1.5594421 1.2728295
## 21    21 0.9170064 1.2758825 1.0964444
## 22    22 0.8822857 1.3908458 1.1365657
```

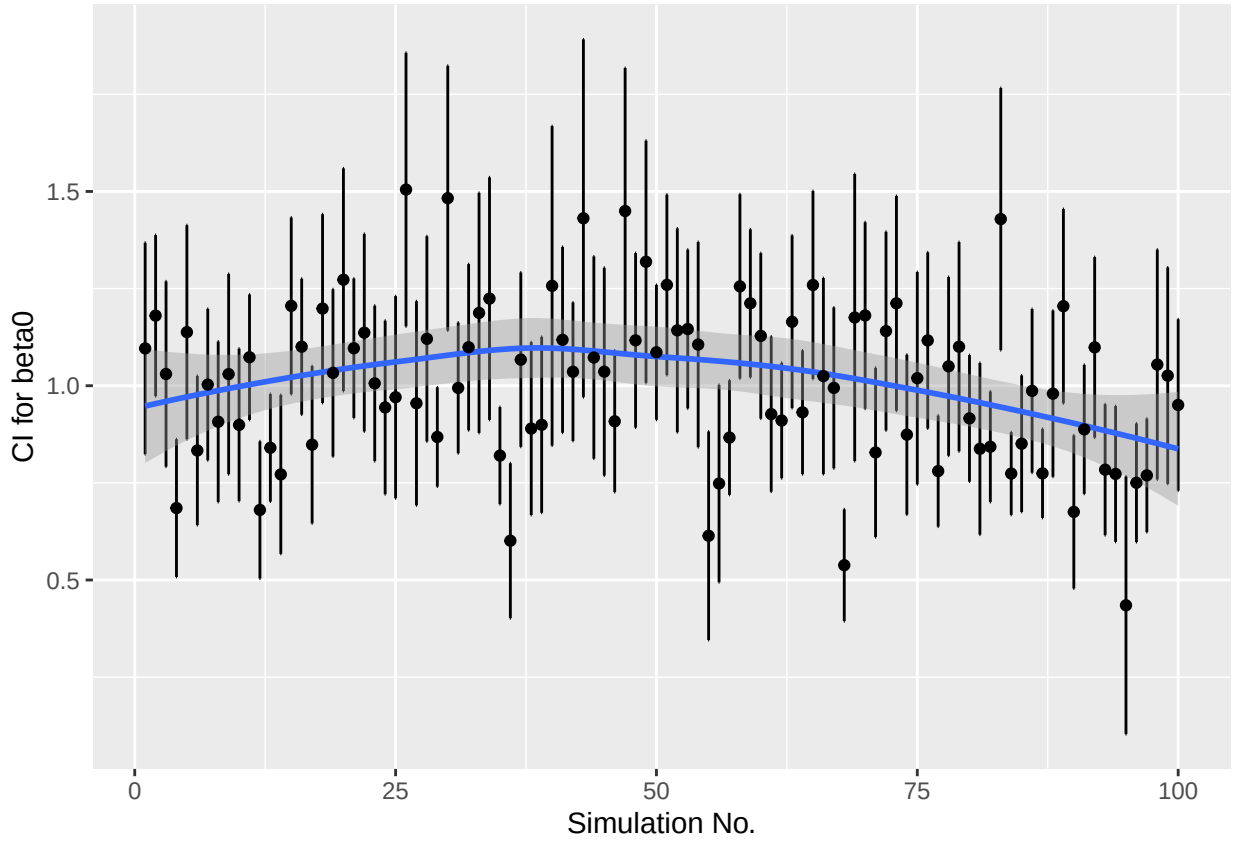
```
## 23   23 0.8060070 1.2060705 1.0060388
## 24   24 0.7207413 1.1674630 0.9441021
## 25   25 0.7101487 1.2306682 0.9704085
## 26   26 1.1527171 1.8569324 1.5048247
## 27   27 0.6920891 1.2176512 0.9548702
## 28   28 0.8565414 1.3847752 1.1206583
## 29   29 0.7404189 0.9958708 0.8681449
## 30   30 1.1419725 1.8234962 1.4827344
## 31   31 0.8260313 1.1625754 0.9943034
## 32   32 0.8852867 1.3130131 1.0991499
## 33   33 0.8784572 1.4964915 1.1874744
## 34   34 0.9121457 1.5364021 1.2242739
## 35   35 0.6951879 0.9451566 0.8201723
## 36   36 0.4017284 0.8004843 0.6011064
## 37   37 0.8427633 1.2915239 1.0671436
## 38   38 0.6672341 1.1122753 0.8897547
## 39   39 0.6730469 1.1258425 0.8994447
## 40   40 0.8459291 1.6682171 1.2570731
## 41   41 0.8789367 1.3573212 1.1181290
## 42   42 0.8578702 1.2142994 1.0360848
## 43   43 0.9705010 1.8914492 1.4309751
## 44   44 0.8123819 1.3328019 1.0725919
## 45   45 0.7687506 1.3036540 1.0362023
## 46   46 0.7262551 1.0905804 0.9084178
## 47   47 1.0816233 1.8176884 1.4496559
## 48   48 0.8921719 1.3414330 1.1168024
## 49   49 1.0073248 1.6307752 1.3190500
## 50   50 0.9123624 1.2597252 1.0860438
## 51   51 1.0270810 1.4920312 1.2595561
## 52   52 0.8802151 1.4048421 1.1425286
## 53   53 0.9415605 1.3502573 1.1459089
## 54   54 0.8419025 1.3699659 1.1059342
## 55   55 0.3454485 0.8820664 0.6137575
## 56   56 0.4942964 1.0022913 0.7482939
## 57   57 0.7191915 1.0140153 0.8666034
## 58   58 1.0185129 1.4929045 1.2557087
## 59   59 1.0213969 1.4026125 1.2120047
## 60   60 0.9151592 1.3413447 1.1282519
## 61   61 0.7272035 1.1266204 0.9269119
## 62   62 0.7618980 1.0589963 0.9104471
## 63   63 0.9423286 1.3867996 1.1645641
## 64   64 0.7722096 1.0908793 0.9315444
## 65   65 1.0172690 1.5012669 1.2592679
## 66   66 0.7724096 1.2772002 1.0248049
## 67   67 0.7873256 1.2014445 0.9943850
## 68   68 0.3941256 0.6820193 0.5380724
## 69   69 0.8061556 1.5449569 1.1755562
## 70   70 0.9407626 1.4208198 1.1807912
## 71   71 0.6102951 1.0462122 0.8282536
## 72   72 0.8849298 1.3962613 1.1405956
## 73   73 0.9356363 1.4885085 1.2120724
## 74   74 0.6681796 1.0799913 0.8740855
## 75   75 0.7460441 1.2927697 1.0194069
## 76   76 0.8898029 1.3432877 1.1165453
```

```
## 77   77 0.6373880 0.9233249 0.7803565
## 78   78 0.8199432 1.2799059 1.0499245
## 79   79 0.8311323 1.3697278 1.1004301
## 80   80 0.7532059 1.0787923 0.9159991
## 81   81 0.6167863 1.0587051 0.8377457
## 82   82 0.7005851 0.9852622 0.8429236
## 83   83 1.0916246 1.7662365 1.4289306
## 84   84 0.6675464 0.8804199 0.7739832
## 85   85 0.6751078 1.0264622 0.8507850
## 86   86 0.7761064 1.1975113 0.9868089
## 87   87 0.6597712 0.8889920 0.7743816
## 88   88 0.7655570 1.1935937 0.9795754
## 89   89 0.9546398 1.4549062 1.2047730
## 90   90 0.4781090 0.8724777 0.6752934
## 91   91 0.7216811 1.0537243 0.8877027
## 92   92 0.8663066 1.3310995 1.0987031
## 93   93 0.6153435 0.9524776 0.7839105
## 94   94 0.5981365 0.9477171 0.7729268
## 95   95 0.1032753 0.7665035 0.4348894
## 96   96 0.5975504 0.9029523 0.7502514
## 97   97 0.6230865 0.9154663 0.7692764
## 98   98 0.7586219 1.3503255 1.0544737
## 99   99 0.7473857 1.3044954 1.0259405
## 100 100 0.7298655 1.1711721 0.9505188
```

```
# Plot

ggplot(dt2SE, aes(x=dt2SE$id, y=dt2SE$coefhat, group=1)) +
  geom_errorbar(aes(ymin=dt2SE$lower, ymax=dt2SE$upper, width=.1)) +
  geom_smooth() +
  geom_point() +
  xlab("Simulation No.") + ylab("CI for beta0")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




```
percent_intervals_with_true_coef <- apply(cbind(dt2SE[,2:3],rep(realparam[2],N)), FUN=interval_contains_

cat("% Coverage Probability for coefficient of 'x' : ", sum(percent_intervals_with_true_coef)/N*100, "%"
```

```
## % Coverage Probability for coefficient of 'x' :  72 %
```

REPEAT Part (c): Wald Test against H_0: $\beta_0 = 1$ (true value)

```
cat("\n Repeating Wald-Test for (Y2, X2)\n")
```

```
##
##  Repeating Wald-Test for (Y2, X2)
```

```
for(i in 1:N)
  WT[i,] <- Wtest(L0, CFs[i,], VHs[i,,], realparam[2])

WTReject <- which(WT[,"p-value"]<0.05)

cat("\n\n No. of Rejections of Ho: ", length(WTReject))
```

```
##
##
##  No. of Rejections of Ho:  29
```

- (e) What is happening here? Give a brief description of what you observe. How does this relate to a 'real world' situation of collecting data?

We can observe that sample #2 has a bias of repeating with higher probability the pairs (X,Y) where X is greater than 1. This changes the distribution of X2's. Besides this, the **glm()** fitted model provide estimates

with greater SEs in the second simulation when compared with the first one resulting in a poor adjustment, this can be seen also through the difference of the AICs.

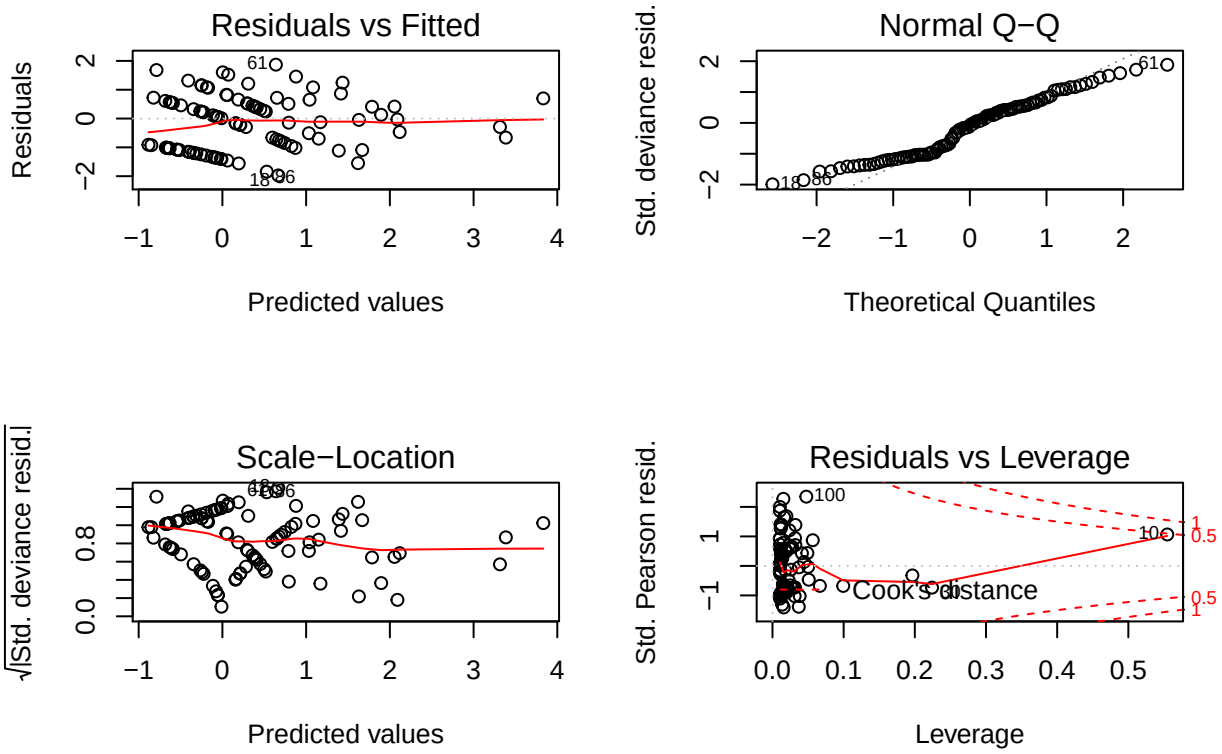
We can observe that the range of 95% CIs for the real value of β_0 is greater on data-set #2 when compared with those obtained in the first simulation. The coverage probability decreased from 96% in the first simulation to 72% in second simuation which denotes less accurate adjustment in the second data-set.

The hipotesis of normality of explanatory variables seems to be violated, with impact when fitting the models in the 2nd. simulation, as well. See examples in 'best' and 'worst' case on each simulation (below)

```
# Model 1 (type 1) - Best Fit in simulation
par(mfrow = c(2, 2))
summary(bestfit1)
```

```
##
## Call:
## glm(formula = Y[i, ] ~ X[i, ] + (XQd[i, ] <- X[i, ]^2), family = poisson())
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.97484  -1.02233  -0.08688   0.53124   1.86958
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.39350    0.09041   4.353 1.35e-05 ***
## X[i, ]                 1.02448    0.09620  10.650  < 2e-16 ***
## XQd[i, ] <- X[i, ]^2  0.20521    0.05108   4.017 5.89e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 584.217  on 99  degrees of freedom
## Residual deviance:  88.738  on 97  degrees of freedom
## AIC: 289.84
##
## Number of Fisher Scoring iterations: 5
```

```
plot(bestfit1)
```



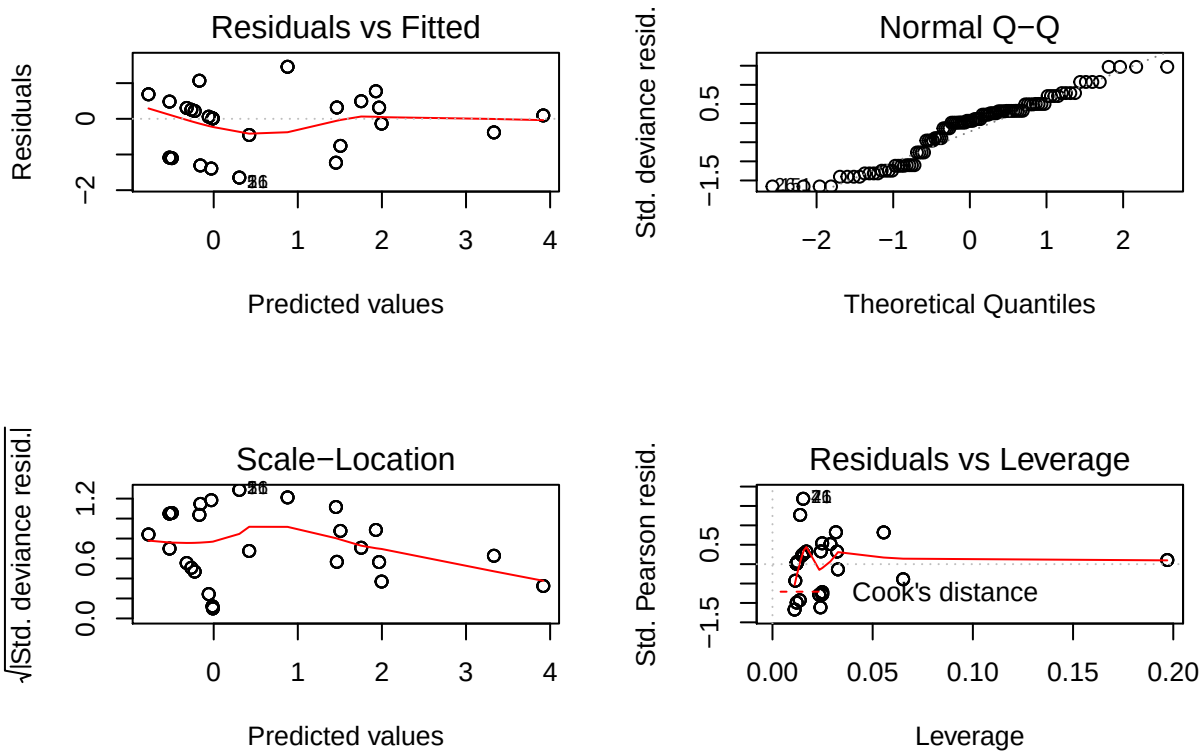

```
par(mfrow = c(1, 1))
```

```
# Model 2 (type 2) - Best Fit in simulation
par(mfrow = c(2, 2))
summary(bestfit2)
```

```
##
## Call:
## glm(formula = Y2[i, ] ~ X2[i, ] + (XQd[i, ] <- X2[i, ]^2), family = poisson())
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.64690  -0.75838   0.05898   0.31730   1.45796
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             0.16740    0.09671   1.731   0.0834 .
## X2[i, ]                 1.02480    0.12620   8.121 4.64e-16 ***
## XQd[i, ] <- X2[i, ]^2   0.26420    0.05326   4.960 7.04e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1168.548  on 99  degrees of freedom
## Residual deviance:   65.787  on 97  degrees of freedom
## AIC: 310.44
```

```
##
## Number of Fisher Scoring iterations: 5
```

```
plot(bestfit2)
```



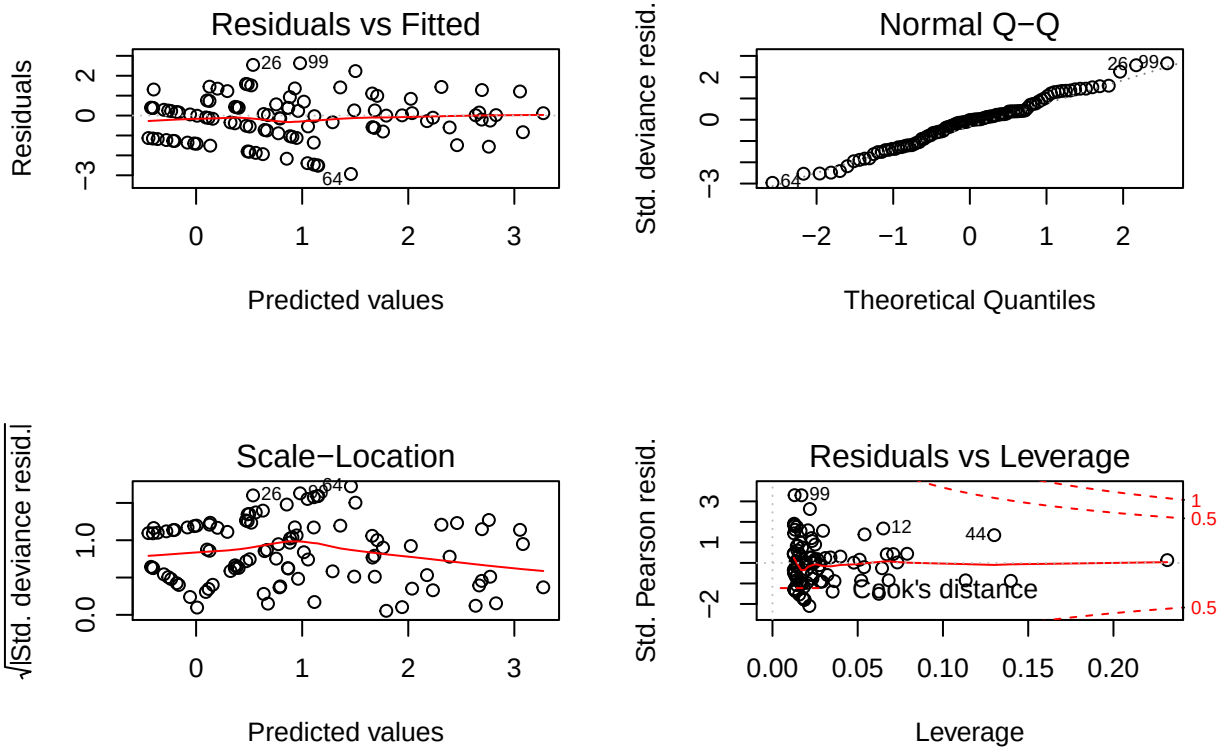

```
par(mfrow = c(1, 1))
```

```
# Model 1 (type 1) - Worst Fit in simulation
par(mfrow = c(2, 2))
summary(worstfit1)
```

```
##
## Call:
## glm(formula = Y[i, ] ~ X[i, ] + (XQd[i, ] <- X[i, ]^2), family = poisson())
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.93413  -1.04608  -0.01934   0.40097   2.63035
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.52724    0.08736   6.035 1.59e-09 ***
## X[i, ]                 0.94695    0.10059   9.414  < 2e-16 ***
## XQd[i, ] <- X[i, ]^2   0.22913    0.06221   3.683  0.00023 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 583.26  on 99  degrees of freedom
## Residual deviance: 133.50  on 97  degrees of freedom
## AIC: 386.12
##
## Number of Fisher Scoring iterations: 5
```

```
plot(worstfit1)
```



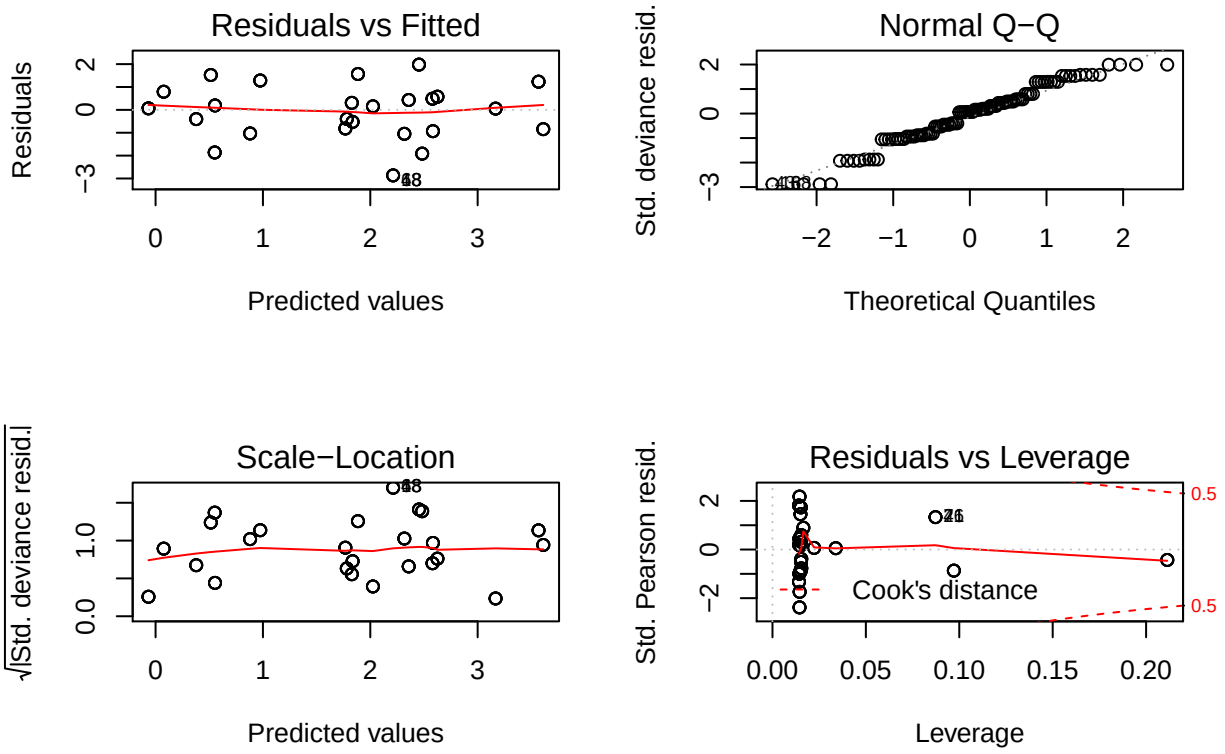

```
par(mfrow = c(1, 1))
```

```
# Model 2 (type 2) - Worst Fit in simulation
par(mfrow = c(2, 2))
summary(worstfit2)
```

```
##
## Call:
## glm(formula = Y2[i, ] ~ X2[i, ] + (XQd[i, ] <- X2[i, ]^2), family = poisson())
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.86363  -0.84335   0.06551   0.57472   1.97811
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            0.45137    0.09500   4.751 2.02e-06 ***
```

```
## X2[i, ]                0.84292    0.07117  11.844  < 2e-16 ***
## XQd[i, ] <- X2[i, ]^2  0.31988    0.03235   9.889  < 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: 909.77  on 99  degrees of freedom
## Residual deviance: 134.90  on 97  degrees of freedom
## AIC: 495.59
##
## Number of Fisher Scoring iterations: 4
```

```
plot(worstfit2)
```




```
par(mfrow = c(1, 1))
```

Question 4 - Opioid mortality in US

The following questions relate to the opioids dataset, which you can find in the data folder of the applied_stats repo. It's an RDS file, which you can read in using read_rds from the tidyverse. There is also a opioids_codebook.txt file which explains each of the variables in the dataset.

The data contains deaths due to opioids by US from 2008 to 2017. In addition, there are population counts and a few other variables of interest. The goal is to explore trends and patterns in opioid deaths over time and across geography. The outcome of interest is deaths.

Please make sure to clearly explain any findings or observations you make, rather than just handing in code and output. You will be assessed not only on the code but also on how you communicate your findings with a combination of writing and analysis.

```
## # A tibble: 6 x 9
##    year state abbrev total_pop deaths expected_deaths prop_white
##   <dbl> <fct> <chr>      <dbl>  <int>           <dbl>      <dbl>
## 1  2008 Alab... AL       4718206    185             301      0.714
## 2  2008 Alas... AK        687455     88              46      0.719
## 3  2008 Ariz... AZ       6280362    494             389      0.865
## 4  2008 Arka... AR       2874554    197             180      0.816
## 5  2008 Cali... CA      36604337   1801            2379      0.765
## 6  2008 Colo... CO       4889730    355             325      0.903
## # ... with 2 more variables: prescription_rate <dbl>, unemp <dbl>
```

(a) Perform some exploratory data analysis (EDA) using this dataset, and briefly summarize in words, tables and charts your main observations. You may use whatever tools or packages you wish. You may want to explore the geofacet package, which plots US state facets in the correct geographic orientation.

Exploratory Data Analysis (EDA)

Preliminarly we will investigate some relations between variables to identify particular behaviours, possible correlations that can provide insights when studying the target variable, i.e., **'deaths'**.

Table 1: Data summary

Name	dtopi
Number of rows	510
Number of columns	9
Column type frequency:	
character	1
factor	1
numeric	7
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
abbrev	0	1	2	2	0	51	0

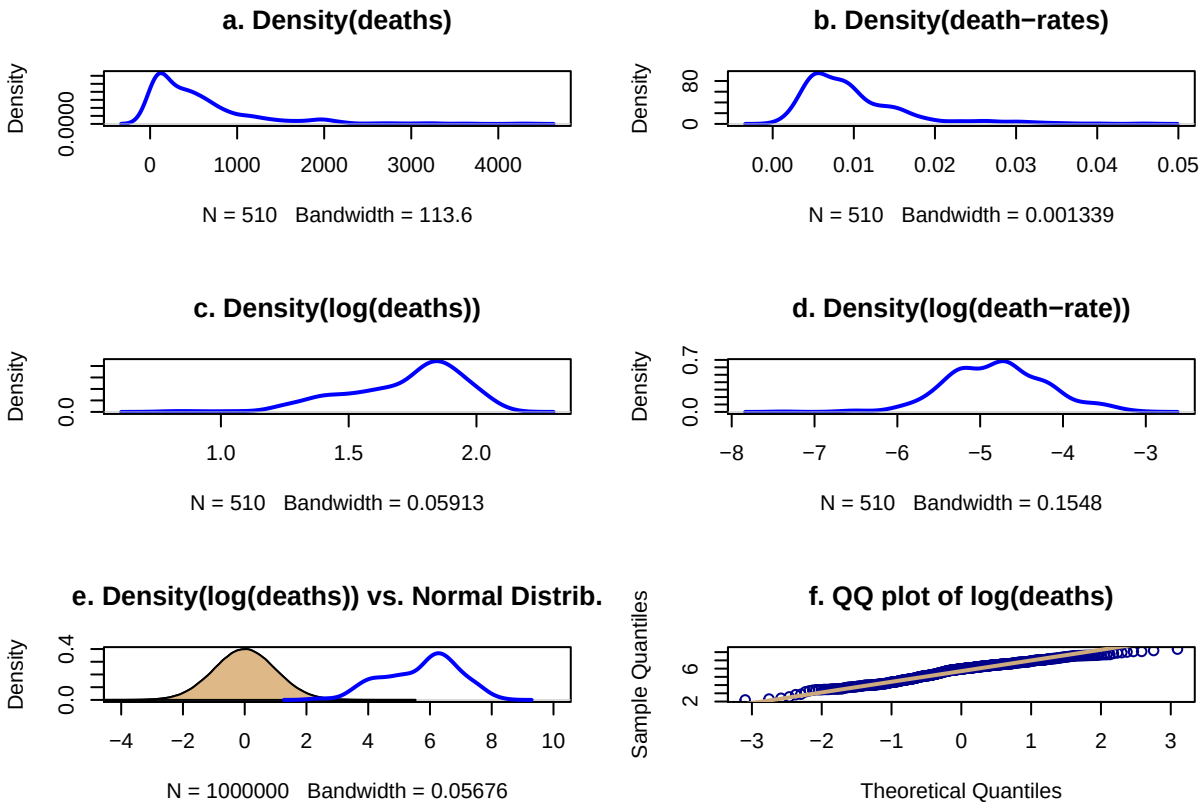
Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
state	0	1	FALSE	51	Ill: 10, Ala: 10, Ala: 10, Ari: 10

Variable type: numeric

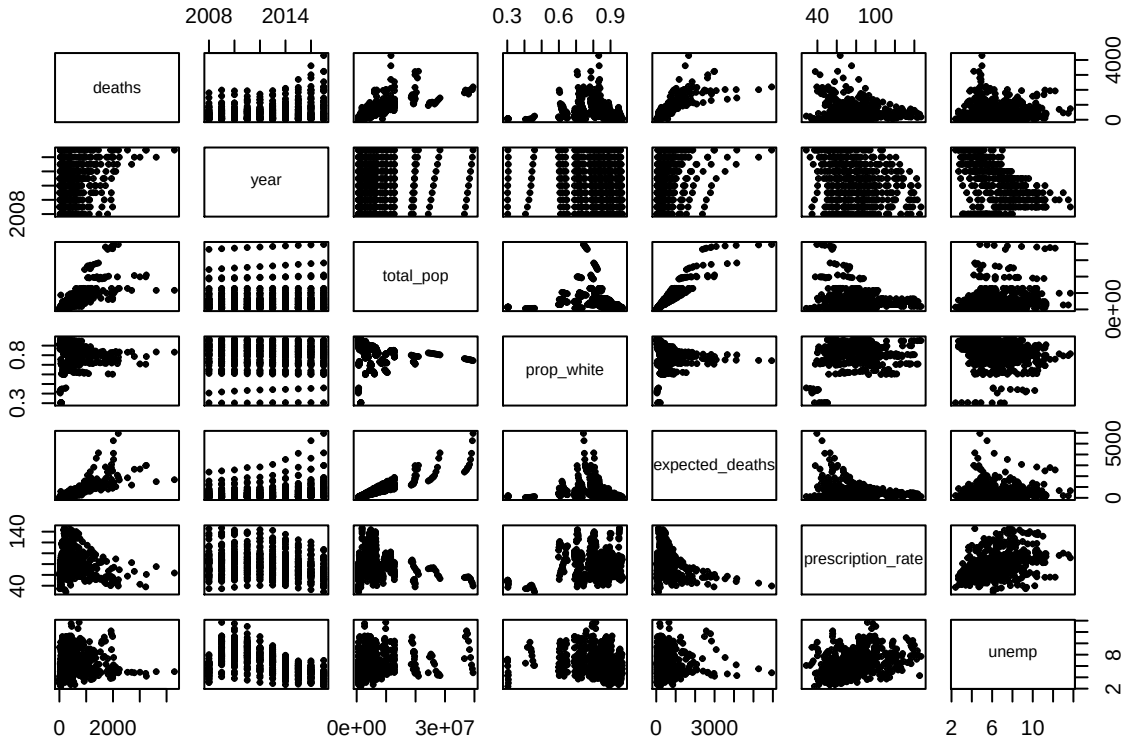
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	
year	0	1	2012.50	2.88	2008.0	2010.00	2012.50	2015
total_pop	0	1	6178163.63	6953751.07	546043.0	1656999.50	4391397.00	6883129
deaths	0	1	556.17	599.70	9.0	124.25	390.50	713
expected_deaths	0	1	556.17	691.61	35.0	138.25	360.50	670
prop_white	0	1	0.81	0.13	0.3	0.73	0.84	0
prescription_rate	0	1	79.60	23.40	28.5	62.30	76.70	92
unemp	0	1	6.46	2.19	2.4	4.80	6.25	7

Initially we compare the density of **no. of deaths** to identify its probable distribution function.



From the graphs above we can verify the distribution of the variable of interest is clearly not a normal but there are simmilarities with Poisson Process or Gamma distributions. The **log(no. of deaths)** seems to be a bi-modal distribution, with approximate composition of two normal distributions (mixture).

This behaviour might suggest that a Poisson-family for `glm()` can be a good choice to investigate the relations between variables and as a start to identify a robust model.

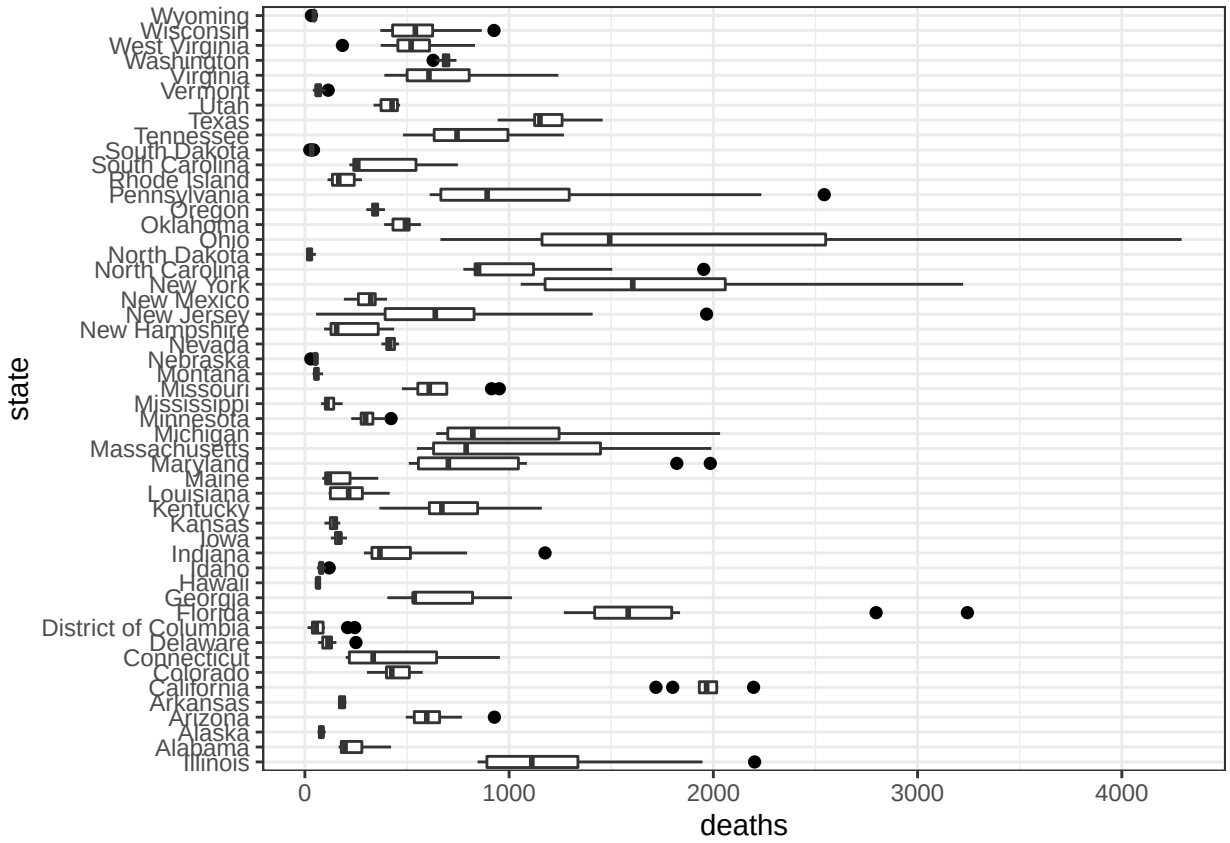


Investigating the relations between covariates through the pair-plot, we visually identify possible tendencies and correlations between the variable of interest (i.e., No. of deaths) and other variables in an overall, i.e., not considering the 'state' sub-classification.

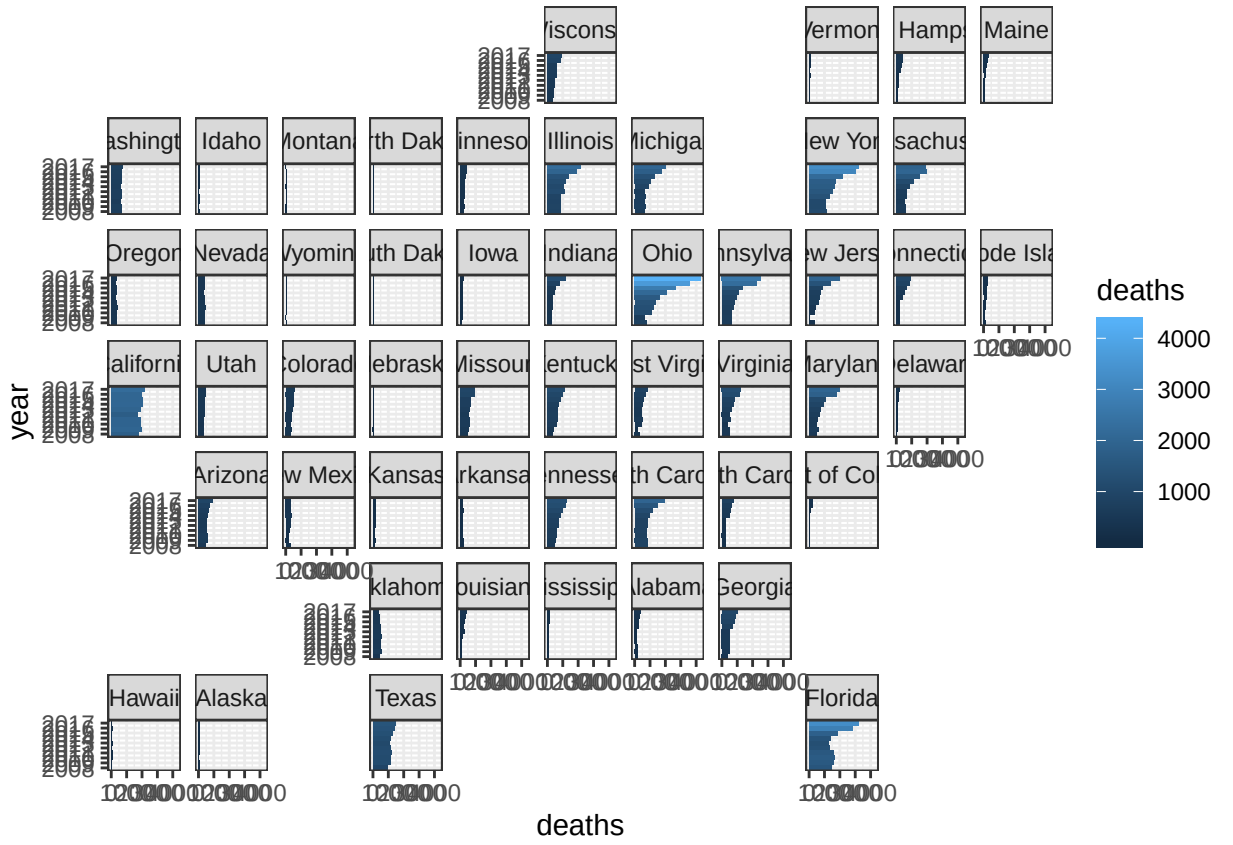
In this graph we can identify some interesting patterns:

- The growth of 'no. of deaths' as 'year' increases;
- Positive correlation between 'no. of deaths' and 'total_pop' suggesting the more total population, we can expect to observe a higher number of deaths;
- Slight higher concentration of 'no. of deaths' with where the proportion of white people is high;
- Decrease of 'no. of deaths' as 'prescription_rate' grows, demonstrating a negative correlation between these variables. This suggests that we can expect less deaths by overdose where people make use of opioids under a prescription;
- Higher concentration of deaths when observed higher unemployment rates and then decreasing as unemployment also decreases.

As this information is another sub-classification of the data, we can investigate which influence 'state' variable plays over the distribution of 'no.of deaths'.

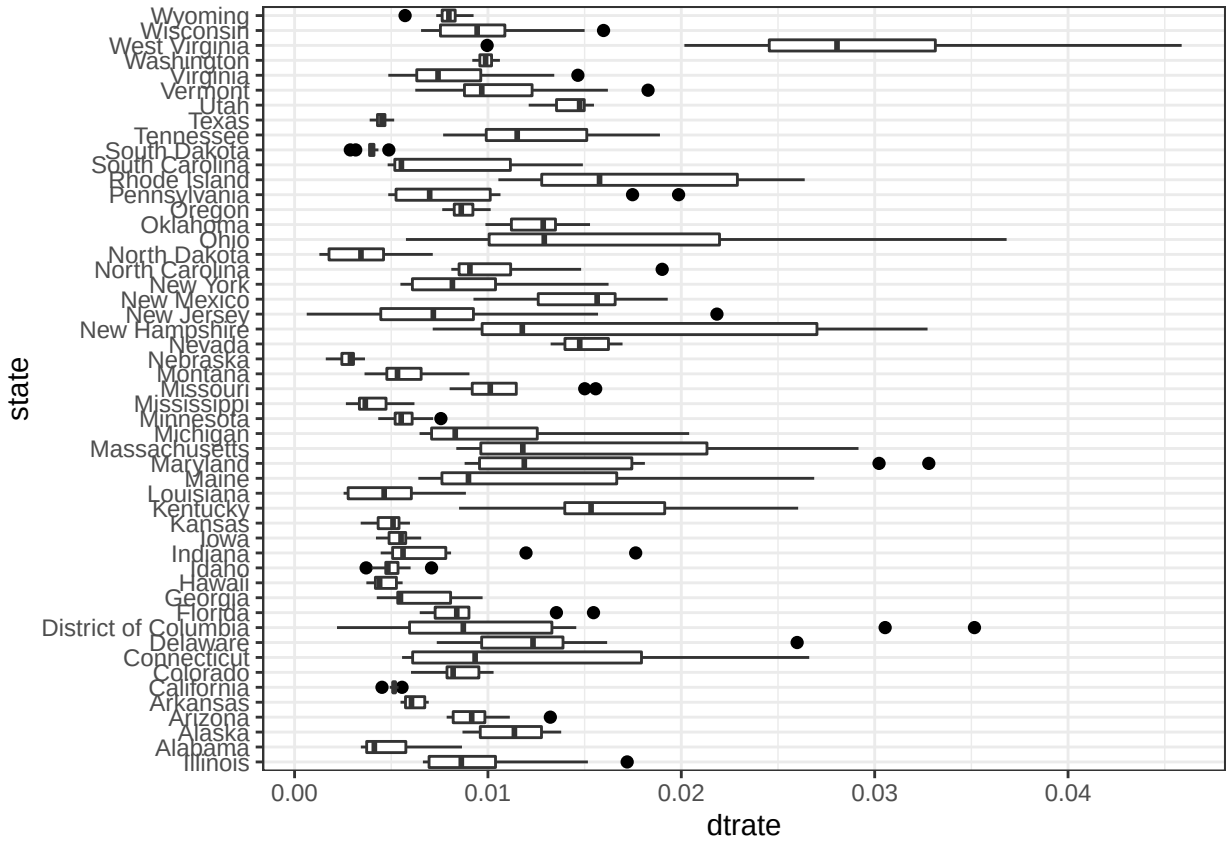


Through the boxplot of 'no. of deaths' per 'state' we can visualize different behaviours on each state. This can be also visualized through the US map view of 'no. of deaths' per year. The charts shows the significant difference between the distributions among the states in US.

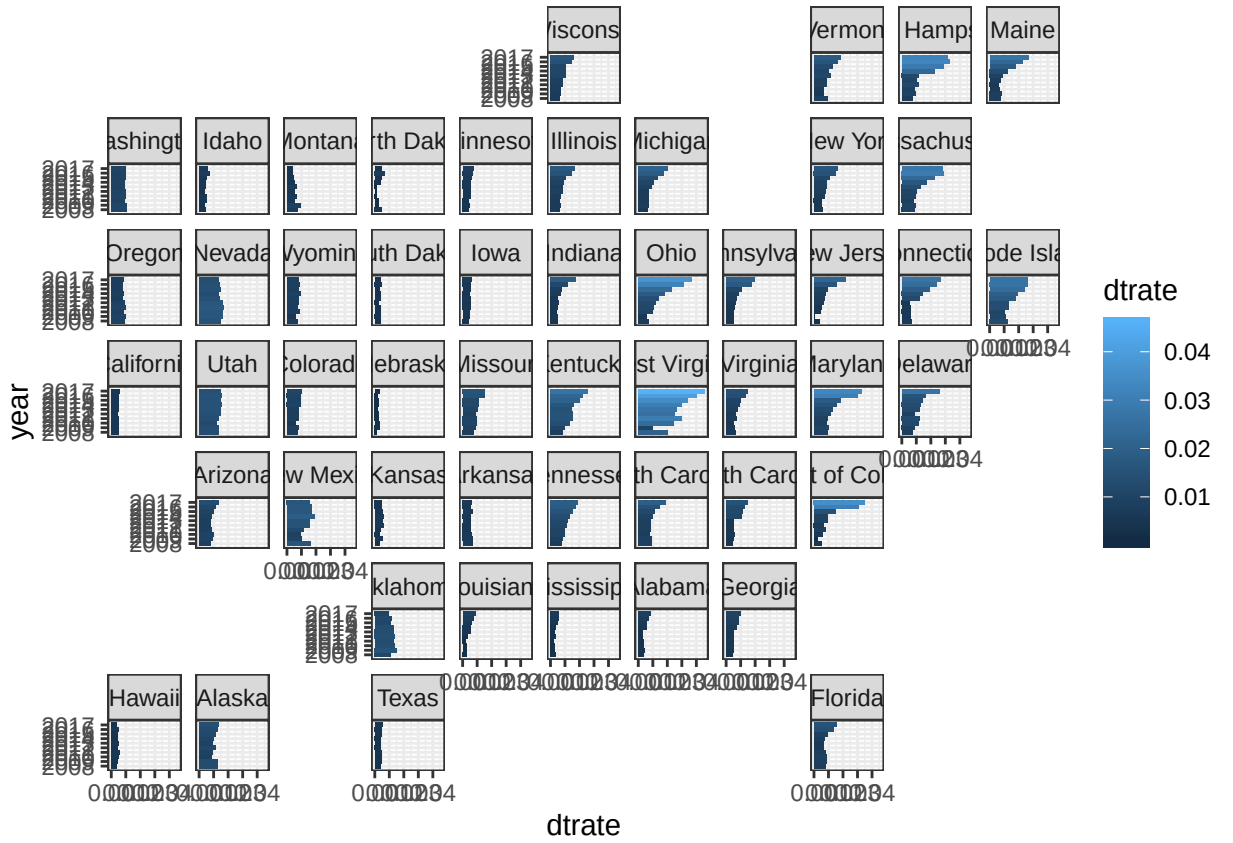


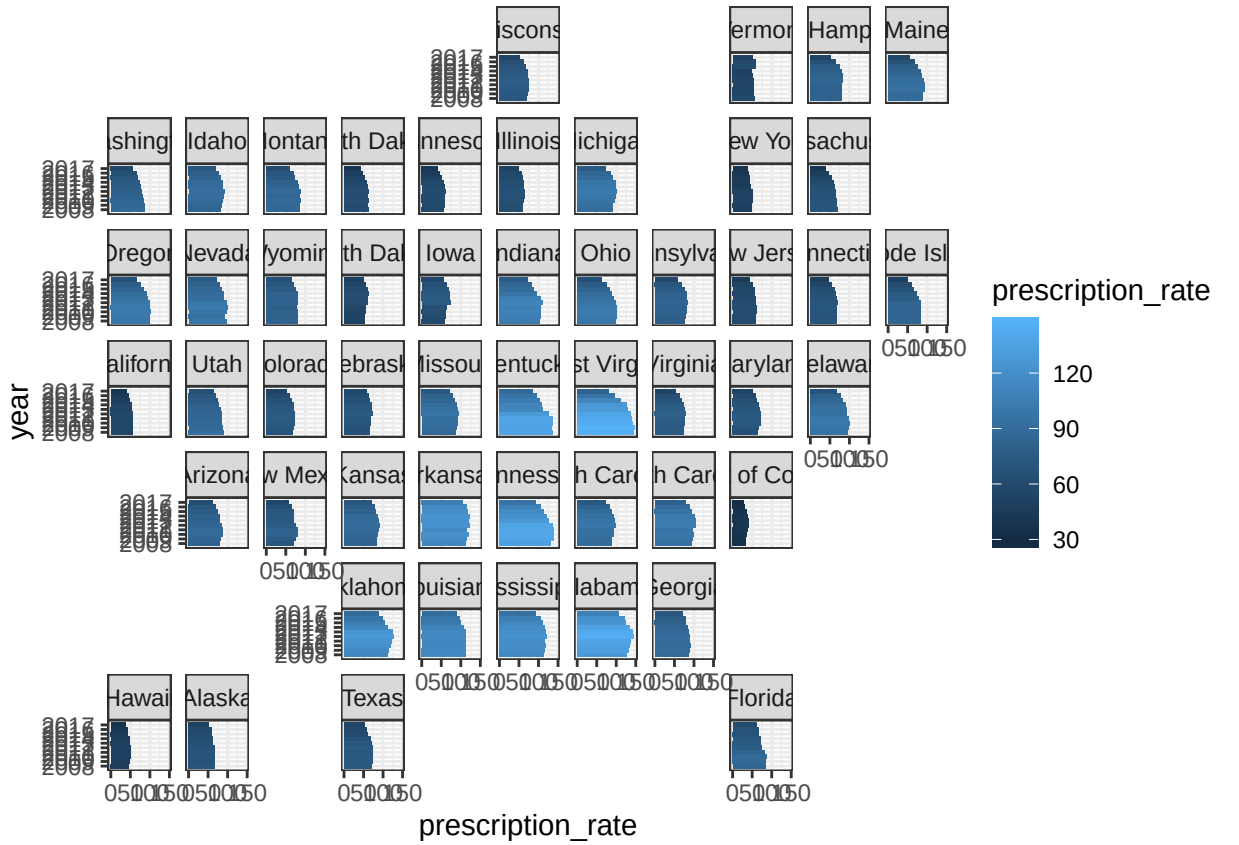
By investigating the boxplot of 'no. of deaths' per state we noticed a clear difference on states of **Ohio**, **New York** and **Pensilvania** when compared with other states.

Now comparing the relative 'no. of deaths' in relation to its population - i.e., comparing the **'death-rate'** of each state, this picture changes significantly as we can see in the next plots.



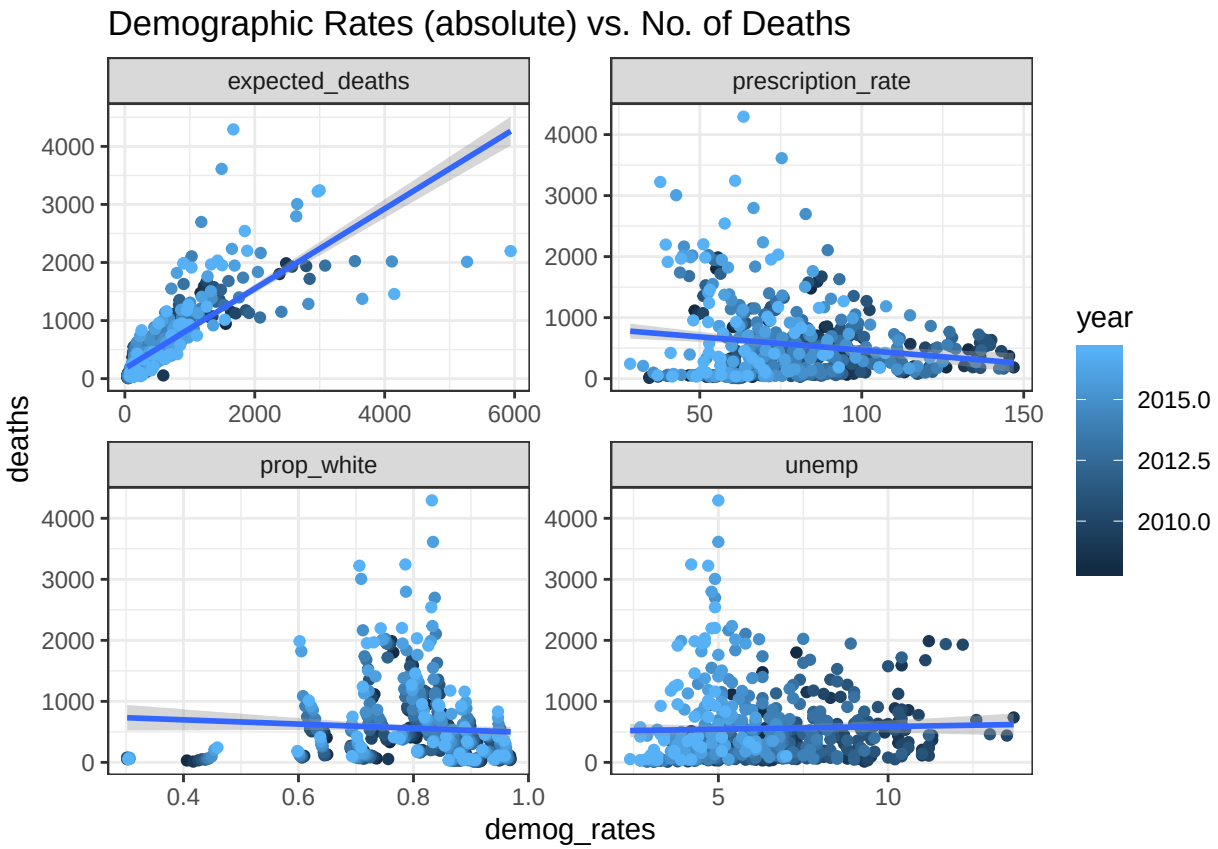
The 'death-rate' of **West Virginia** is the largest one, followed by **Ohio** and **New Hampshire**. We can visualize such different shape in the US map using the modified variable 'death-rate'.


year
dtrate
dtrate
0.04
0.03
0.02
0.01
Hawaii
Alaska
Texas
Florida



Some comments arise in this context:

- Apparently the states from the East-side demonstrated increase of **death-rates** (deaths over state population) during the period analysed while, on the other hand, Central and West-side states remained slightly stable in the period.
- States such as West Virginia, Ohio, Maryland, New Hampshire and Massachussets are among the highest variation from 2008-2017;
- **Prescription rate** seems to play an interesting role in those high-mortality east-side states, where the higher **prescription-rate** seems to be negatively correlated with the **no. of deaths**. But this behaviour seems to be different in those states with higher death-rates (e.g., West Virginia) and may be related to other factor.



A closer view over the interest variable (deaths) against the explanatory covariates shows some relations between **deaths** and **prescription_rate** suggesting that an increasing prescription rate, may contribute for a decreasing no. of deaths registered.

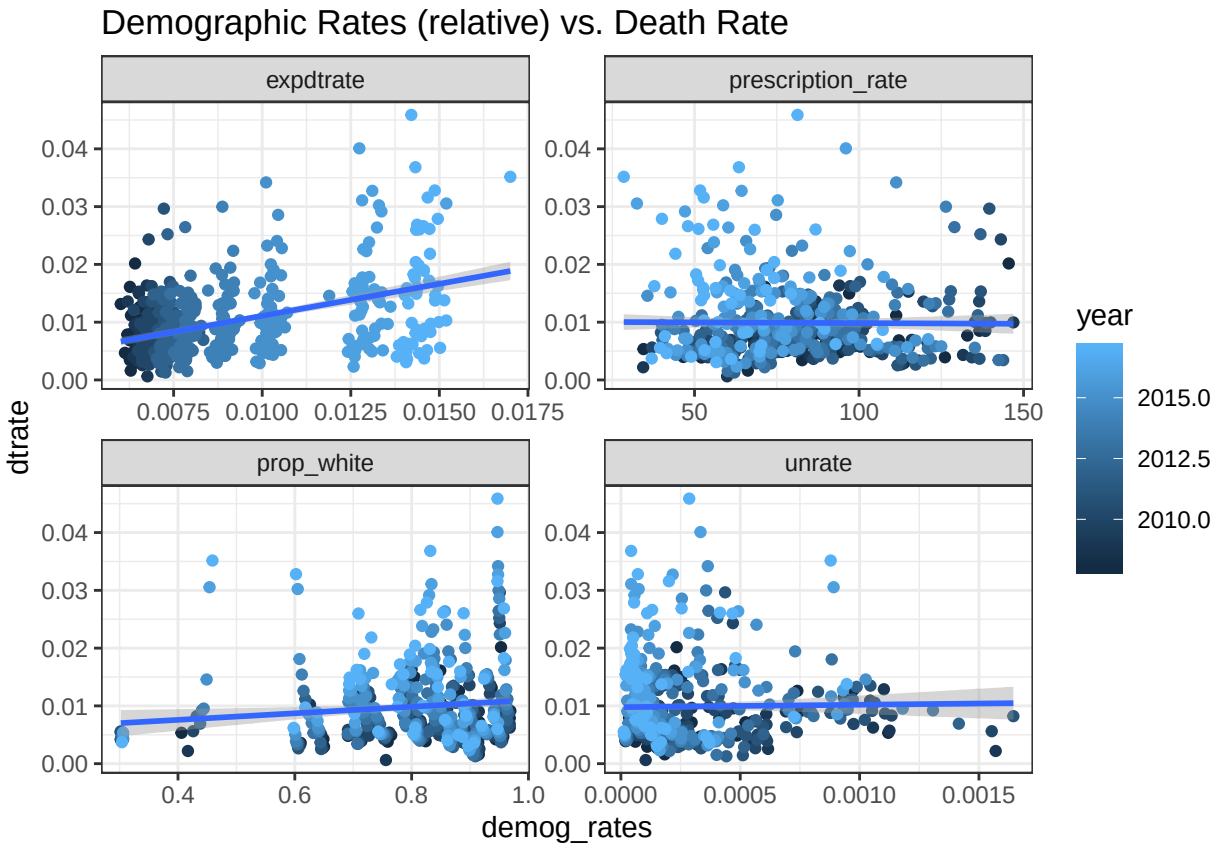
No. of Deaths seems to be negatively correlated with **prescription rate**, suggesting that higher prescription rate is followed by a reduced no. of deaths. Besides, the pairs of scatterplots show that places with greater proportion of white people we observe reduced no. of deaths. On the other hand, the expected no. of deaths is positively correlated with **no. of deaths**.

The same behaviour can be noted when using **death-rate** (i.e., 'no.of deaths'/'total_pop') against those variables.

```
##                       deaths  total_pop expected_deaths  prop_white
## deaths             1.00000000  0.7372172       0.7940006 -0.07503019
## total_pop          0.73721722  1.0000000       0.9269393 -0.11796177
## expected_deaths    0.79400065  0.9269393       1.0000000 -0.12170262
## prop_white        -0.07503019 -0.1179618      -0.1217026  1.00000000
## prescription_rate -0.16966356 -0.1783913      -0.2412520  0.13820934
## unemp              0.03197441  0.2067809       0.0268111 -0.16851047
##                   prescription_rate       unemp
## deaths                   -0.1696636  0.03197441
## total_pop                -0.1783913  0.20678087
## expected_deaths          -0.2412520  0.02681110
## prop_white                0.1382093 -0.16851047
## prescription_rate         1.0000000  0.35091378
## unemp                     0.3509138  1.00000000
```

Correlations between **deaths** and **total_pop/expected_deaths** are higher than between the others so

these variables might be probable candidates to have in the model. On the other hand, variables **prop_white** and **unemp** have low correlation with 'no. of deaths', suggesting their influence might be lower.



(b) Run a Poisson regression using deaths as the outcome and tot_pop as the offset. (remember to log the offset). Include the state variable as a factor and change the reference category to be Illinois (you can do this using the relevel function). Investigate which variables to include, justifying based on your EDA in part a). Interpret your findings, including visualizations where appropriate. Include an analysis of which states, after accounting for other variables in the model, have the highest opioid mortality.

First adjusting factor variables and relevel of variable 'state' to "Illinois".

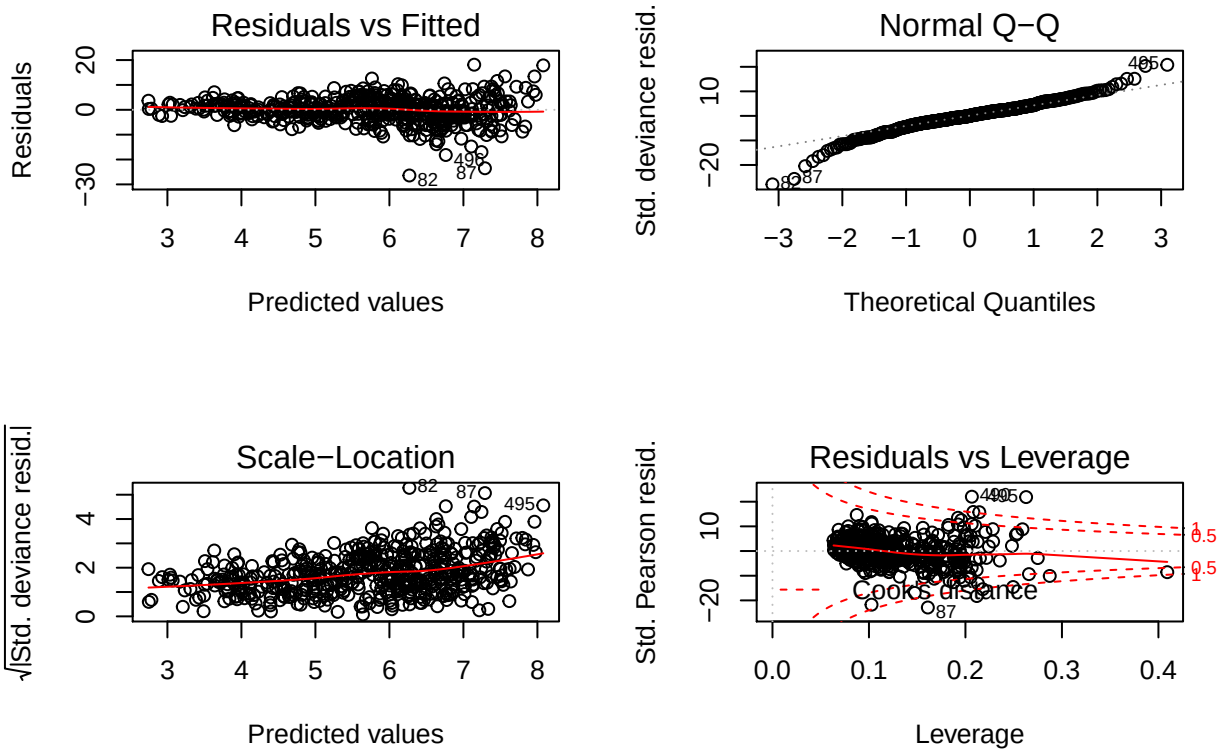
Running the basic model using '**tot_pop** as offset. Initially we will include all variables and check which ones are significant to predict mortality by using opioids.

```
##
## Call:
## glm(formula = deaths ~ year + state + expected_deaths + prop_white +
##     prescription_rate + unemp, family = poisson(), data = dt,
##     offset = log(total_pop))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -26.4251   -2.4936    0.0475    2.7709   18.1289
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -1.057e+01  5.656e-01 -18.689  < 2e-16 ***
```

```
## year2009                  -1.498e-02  1.468e-02  -1.020 0.307748
## year2010                   7.942e-03  1.548e-02   0.513 0.607863
## year2011                   1.039e-01  1.398e-02   7.434 1.05e-13 ***
## year2012                   1.328e-01  1.283e-02  10.348  < 2e-16 ***
## year2013                   2.353e-01  1.266e-02  18.595  < 2e-16 ***
## year2014                   4.118e-01  1.303e-02  31.608  < 2e-16 ***
## year2015                   6.026e-01  1.493e-02  40.361  < 2e-16 ***
## year2016                   9.135e-01  1.733e-02  52.720  < 2e-16 ***
## year2017                   1.083e+00  2.087e-02  51.896  < 2e-16 ***
## stateAlabama              -8.311e-01  6.624e-02 -12.547  < 2e-16 ***
## stateAlaska                2.991e-02  6.905e-02   0.433 0.664916
## stateArizona              -2.395e-01  5.031e-02  -4.761 1.92e-06 ***
## stateArkansas             -7.310e-01  4.161e-02 -17.571  < 2e-16 ***
## stateCalifornia           -1.703e-01  2.985e-02  -5.703 1.17e-08 ***
## stateColorado             -3.758e-01  7.742e-02  -4.854 1.21e-06 ***
## stateConnecticut           7.078e-02  3.305e-02   2.141 0.032237 *
## stateDelaware              1.474e-01  5.821e-02   2.532 0.011343 *
## stateDistrict of Columbia  5.199e-01  2.436e-01   2.134 0.032855 *
## stateFlorida               3.184e-02  1.406e-02   2.265 0.023533 *
## stateGeorgia              -3.068e-01  1.094e-01  -2.805 0.005032 **
## stateHawaii               -3.194e-01  3.422e-01  -0.934 0.350560
## stateIdaho                -1.056e+00  1.217e-01  -8.675  < 2e-16 ***
## stateIndiana              -5.031e-01  6.512e-02  -7.726 1.11e-14 ***
## stateIowa                 -8.929e-01  1.035e-01  -8.625  < 2e-16 ***
## stateKansas               -9.736e-01  7.562e-02 -12.875  < 2e-16 ***
## stateKentucky              1.606e-01  8.291e-02   1.937 0.052744 .
## stateLouisiana            -7.378e-01  1.047e-01  -7.043 1.88e-12 ***
## stateMaine                -1.342e-01  1.262e-01  -1.063 0.287834
## stateMaryland              5.901e-01  1.228e-01   4.805 1.55e-06 ***
## stateMassachusetts         3.543e-01  3.986e-02   8.887  < 2e-16 ***
## stateMichigan             -5.177e-02  2.684e-02  -1.929 0.053780 .
## stateMinnesota            -7.070e-01  6.155e-02 -11.486  < 2e-16 ***
## stateMississippi          -9.128e-01  1.325e-01  -6.888 5.64e-12 ***
## stateMissouri             -9.622e-02  4.780e-02  -2.013 0.044107 *
## stateMontana              -8.814e-01  9.728e-02  -9.061  < 2e-16 ***
## stateNebraska             -1.539e+00  9.389e-02 -16.387  < 2e-16 ***
## stateNevada                1.925e-01  2.365e-02   8.140 3.96e-16 ***
## stateNew Hampshire         2.272e-01  1.179e-01   1.927 0.054000 .
## stateNew Jersey           -1.474e-01  3.434e-02  -4.293 1.76e-05 ***
## stateNew Mexico            1.563e-01  4.364e-02   3.581 0.000342 ***
## stateNew York              1.760e-01  5.042e-02   3.491 0.000482 ***
## stateNorth Carolina        6.799e-02  4.515e-02   1.506 0.132093
## stateNorth Dakota         -1.290e+00  1.016e-01 -12.702  < 2e-16 ***
## stateOhio                  4.282e-01  4.213e-02  10.164  < 2e-16 ***
## stateOklahoma              4.305e-02  3.061e-02   1.407 0.159570
## stateOregon               -4.420e-01  8.438e-02  -5.238 1.62e-07 ***
## statePennsylvania         -1.394e-01  4.060e-02  -3.433 0.000598 ***
## stateRhode Island          2.756e-01  6.160e-02   4.474 7.66e-06 ***
## stateSouth Carolina       -2.685e-01  6.926e-02  -3.877 0.000106 ***
## stateSouth Dakota         -1.152e+00  8.005e-02 -14.390  < 2e-16 ***
## stateTennessee             2.764e-02  3.396e-02   0.814 0.415699
## stateTexas                -5.639e-01  2.477e-02 -22.769  < 2e-16 ***
## stateUtah                  4.777e-02  1.021e-01   0.468 0.639888
## stateVermont              -2.426e-01  1.286e-01  -1.887 0.059174 .
```

```
## stateVirginia             -1.081e-01  4.983e-02  -2.168 0.030123 *
## stateWashington           -1.539e-01  3.668e-02  -4.194 2.74e-05 ***
## stateWest Virginia         6.129e-01  1.202e-01   5.100 3.40e-07 ***
## stateWisconsin            -1.891e-01  7.328e-02  -2.581 0.009846 **
## stateWyoming              -5.921e-01  1.202e-01  -4.925 8.43e-07 ***
## expected_deaths           -1.681e-04  6.370e-06 -26.392  < 2e-16 ***
## prop_white                 1.148e+00  7.003e-01   1.639 0.101240
## prescription_rate          1.819e-03  4.411e-04   4.125 3.71e-05 ***
## unemp                      1.551e-02  3.068e-03   5.055 4.29e-07 ***
## ---
## 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: 12552  on 446  degrees of freedom
## AIC: 16542
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] 16542
```



In this adjustment we have the following model:

$$\mu_{deaths_i} = \alpha + \beta_1 \times factor(year_i) + \beta_2 \times factor(state_i) + \beta_3 \times expected\_deaths_i + \beta_4 \times prop\_white_i + \beta_5 \times prescription\_rate_i \beta_6 \times unemp_i + 1 \times log(total\_pop) \quad (20)$$

Apparently we identified 03 potential outliers (or influent points) in the observed data: [82, 87, 496]:

```
## # A tibble: 3 x 14
##   year  state abbrev total_pop deaths expected_deaths prop_white
##   <fct> <fct> <fct>      <dbl>  <int>           <dbl>      <dbl>
## 1 2009  New ... NJ       8755602     55             592      0.756
## 2 2009  Ohio  OH      11528896    664             760      0.848
## 3 2017  Okla... OK       3930864    388             559      0.776
## # ... with 7 more variables: prescription_rate <dbl>, unemp <dbl>, dtrate <dbl>,
## #   expdtrate <dbl>, unrate <dbl>, difdeath <dbl>, logdeath <dbl>
```

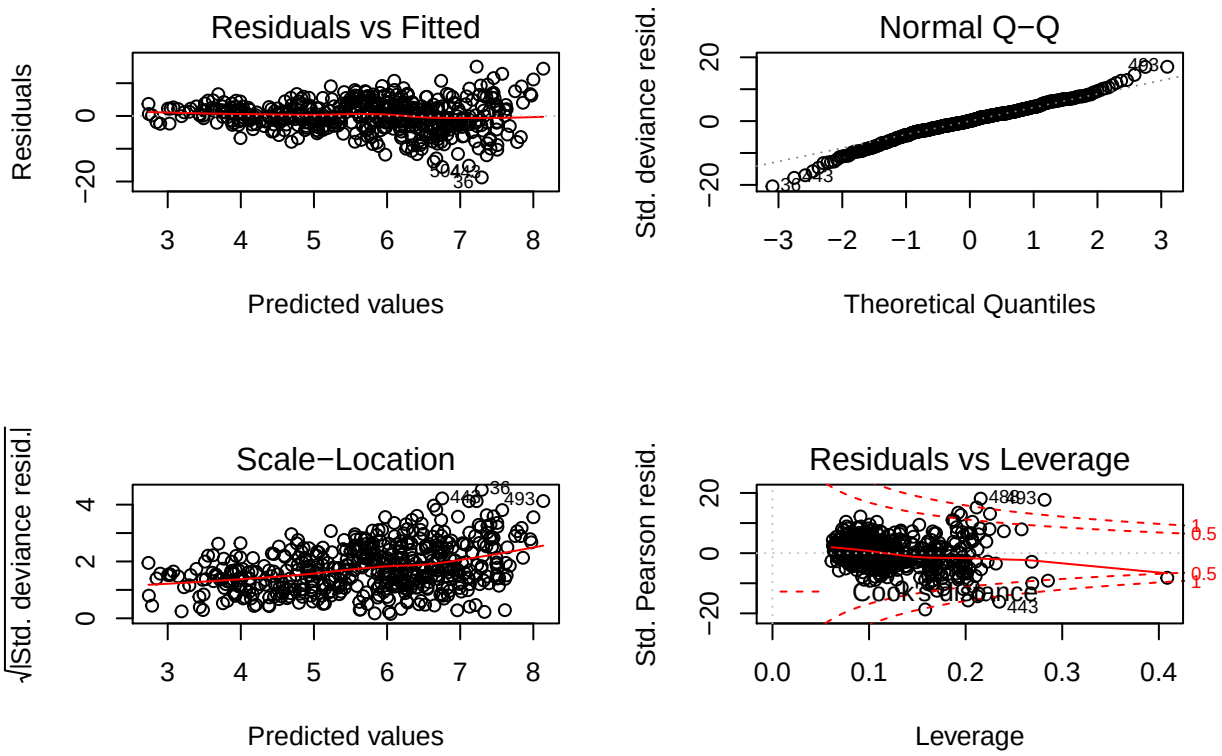
Fom the fitted model **fit1** it can be identified the variable **prop_white** is not significant and then we will remove it from the model. The **factor(year)** and **factor(state)** represent the variables treated as factors and some values are not sigificant, as well, such as $year_i = 2009$, $state_i =$ "*Utah*" among others. This is reflected with coefficient near '0'.

We will then fit another model by removing the potential outliers, supressing variable **prop_white** and maintaining the other parameters unchanged.

```
##
## Call:
## glm(formula = deaths ~ year + state + expected_deaths + prescription_rate +
##     unemp, family = poisson(), data = dt[-outliers, ], offset = log(total_pop))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -18.7586   -2.5587    0.0804    2.7430   15.0457
##
## Coefficients:
##                              Estimate Std. Error  z value Pr(>|z|)
## (Intercept)                -9.747e+00  3.656e-02 -266.583  < 2e-16 ***
## year2009                    1.854e-02  1.466e-02    1.265 0.205984
## year2010                   -3.769e-02  1.549e-02   -2.433 0.014968 *
## year2011                    6.398e-02  1.376e-02    4.651 3.31e-06 ***
## year2012                    9.956e-02  1.203e-02    8.277  < 2e-16 ***
## year2013                    2.072e-01  1.098e-02   18.867  < 2e-16 ***
## year2014                    3.945e-01  9.920e-03   39.764  < 2e-16 ***
## year2015                    5.930e-01  1.063e-02   55.801  < 2e-16 ***
## year2016                    9.072e-01  1.242e-02   73.069  < 2e-16 ***
## year2017                    1.094e+00  1.557e-02   70.280  < 2e-16 ***
## stateAlabama               -9.362e-01  3.883e-02  -24.108  < 2e-16 ***
## stateAlaska                -6.403e-02  3.711e-02   -1.725 0.084476 .
## stateArizona               -1.645e-01  1.826e-02   -9.009  < 2e-16 ***
## stateArkansas              -7.024e-01  3.757e-02  -18.697  < 2e-16 ***
## stateCalifornia            -2.097e-01  2.113e-02   -9.924  < 2e-16 ***
## stateColorado              -2.317e-01  1.983e-02  -11.683  < 2e-16 ***
## stateConnecticut            1.234e-01  1.875e-02    6.583 4.62e-11 ***
## stateDelaware               7.832e-02  3.452e-02    2.269 0.023278 *
## stateDistrict of Columbia   1.292e-01  3.717e-02    3.474 0.000512 ***
## stateFlorida                4.096e-02  1.321e-02    3.101 0.001932 **
## stateGeorgia               -4.888e-01  1.861e-02  -26.269  < 2e-16 ***
```

```
## stateHawaii               -8.395e-01  4.245e-02  -19.775  < 2e-16 ***
## stateIdaho                -8.536e-01  3.945e-02  -21.639  < 2e-16 ***
## stateIndiana              -4.045e-01  2.420e-02  -16.712  < 2e-16 ***
## stateIowa                 -6.989e-01  2.905e-02  -24.058  < 2e-16 ***
## stateKansas               -8.415e-01  3.218e-02  -26.149  < 2e-16 ***
## stateKentucky              2.761e-01  2.951e-02    9.357  < 2e-16 ***
## stateLouisiana            -9.042e-01  3.203e-02  -28.230  < 2e-16 ***
## stateMaine                 8.104e-02  2.953e-02    2.745 0.006055 **
## stateMaryland              4.080e-01  1.580e-02   25.821  < 2e-16 ***
## stateMassachusetts         4.347e-01  1.491e-02   29.144  < 2e-16 ***
## stateMichigan             -3.529e-02  1.937e-02   -1.822 0.068472 .
## stateMinnesota            -5.838e-01  2.225e-02  -26.236  < 2e-16 ***
## stateMississippi          -1.136e+00  3.906e-02  -29.098  < 2e-16 ***
## stateMissouri             -2.019e-02  2.074e-02   -0.973 0.330472
## stateMontana              -7.209e-01  4.514e-02  -15.971  < 2e-16 ***
## stateNebraska             -1.366e+00  4.753e-02  -28.739  < 2e-16 ***
## stateNevada                1.709e-01  2.362e-02    7.232 4.75e-13 ***
## stateNew Hampshire         4.462e-01  2.737e-02   16.306  < 2e-16 ***
## stateNew Jersey           -1.205e-01  1.504e-02   -8.016 1.10e-15 ***
## stateNew Mexico            2.217e-01  2.251e-02    9.846  < 2e-16 ***
## stateNew York              1.116e-01  1.452e-02    7.687 1.50e-14 ***
## stateNorth Carolina       -5.063e-03  1.860e-02   -0.272 0.785442
## stateNorth Dakota         -1.115e+00  6.570e-02  -16.974  < 2e-16 ***
## stateOhio                  5.405e-01  1.683e-02   32.108  < 2e-16 ***
## stateOklahoma              1.703e-01  3.121e-02    5.456 4.87e-08 ***
## stateOregon               -3.136e-01  2.440e-02  -12.851  < 2e-16 ***
## statePennsylvania         -7.052e-02  1.484e-02   -4.753 2.00e-06 ***
## stateRhode Island          3.597e-01  2.676e-02   13.440  < 2e-16 ***
## stateSouth Carolina       -3.826e-01  2.524e-02  -15.157  < 2e-16 ***
## stateSouth Dakota         -1.020e+00  5.786e-02  -17.631  < 2e-16 ***
## stateTennessee             3.687e-02  3.174e-02    1.162 0.245295
## stateTexas                -5.146e-01  1.547e-02  -33.267  < 2e-16 ***
## stateUtah                  2.357e-01  2.349e-02   10.035  < 2e-16 ***
## stateVermont              -7.802e-03  4.158e-02   -0.188 0.851164
## stateVirginia             -1.621e-01  1.760e-02   -9.211  < 2e-16 ***
## stateWashington           -9.682e-02  1.712e-02   -5.656 1.55e-08 ***
## stateWest Virginia         7.928e-01  3.321e-02   23.869  < 2e-16 ***
## stateWisconsin            -5.583e-02  1.828e-02   -3.054 0.002258 **
## stateWyoming              -3.912e-01  5.018e-02   -7.794 6.47e-15 ***
## expected_deaths           -1.694e-04  6.296e-06  -26.905  < 2e-16 ***
## prescription_rate          2.072e-03  4.403e-04    4.705 2.54e-06 ***
## unemp                      2.574e-02  3.023e-03    8.514  < 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: 86160  on 506  degrees of freedom
## Residual deviance: 10749  on 444  degrees of freedom
## AIC: 14715
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] 14715.35
```



The main aspects we can notice in this adjustment are:

- Controlling by (state = "Illinois"), the model will incorporate this info by modelling the coefficients in such a way that states with higher 'no. of deaths' will present higher and *positive* coefficients and, on the other hand, those states with significant low 'no. of deaths'will present *negative* coefficients
- The higher opioid mortality is observed in state of **West Virginia** followed by **Ohio** and **New Hampshire** due to its contribution with positive coefficient among the states in the adjusted model. This agrees with EDA Analysis done in item (a).
- The higher prescription rate is present on the state, the higher mortality is observed;

(c) What's an issue with using population as an offset, given the limited information available in this dataset? (hint: the probability of death varies by age).

As the probability of death changes with age it would be plausible to expect that states with older population would show higher no. of deaths. As 'age' is not present in dataset, we can't put this variable in perspective to investigate this behaviour.

The issue by using **population** as an offset, this will include **log(total_pop)** in the model with coefficient '1' (instead of an estimated one) and then we can expect the variable of interest be weightened differently depending on how large (or how dense) is the population on each state.

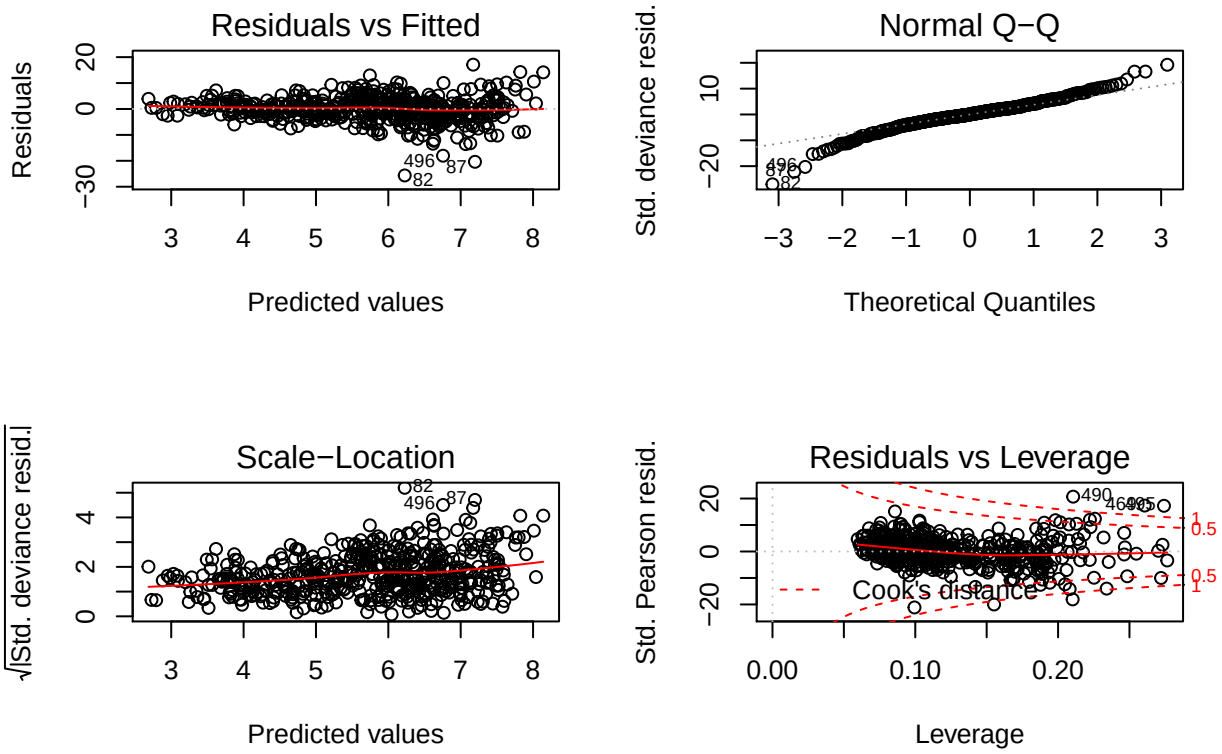
In other words, by doing this, we will just assume that 'the more people on the state' implies 'the more no. of deaths' there, not capturing other nuances that may affect the variable of interest, such as 'age'.

(d) Rerun your Poisson regression using expected_deaths as an offset. How does this change the interpretation of your coefficients?

```
##
## Call:
## glm(formula = deaths ~ year + state + prop_white + prescription_rate +
##     unemp + total_pop, family = poisson(), data = dt, offset = log(expected_deaths))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -25.6492   -2.3572   -0.0122    2.3987   17.0935
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.853e+00  5.784e-01   4.932 8.12e-07 ***
## year2009                   -1.118e-02  1.467e-02  -0.762 0.445788
## year2010                    1.383e-02  1.561e-02   0.886 0.375392
## year2011                    5.067e-02  1.418e-02   3.573 0.000353 ***
## year2012                    8.796e-02  1.312e-02   6.704 2.03e-11 ***
## year2013                    1.217e-01  1.297e-02   9.383  < 2e-16 ***
## year2014                    1.637e-01  1.330e-02  12.309  < 2e-16 ***
## year2015                    2.023e-01  1.509e-02  13.408  < 2e-16 ***
## year2016                    2.333e-01  1.712e-02  13.628  < 2e-16 ***
## year2017                    2.645e-01  2.039e-02  12.968  < 2e-16 ***
## stateAlabama               -2.759e+00  8.152e-02 -33.846  < 2e-16 ***
## stateAlaska                -3.059e+00  9.829e-02 -31.119  < 2e-16 ***
## stateArizona               -1.650e+00  5.834e-02 -28.284  < 2e-16 ***
## stateArkansas              -3.052e+00  6.742e-02 -45.273  < 2e-16 ***
## stateCalifornia             6.113e+00  1.381e-01  44.270  < 2e-16 ***
## stateColorado              -2.195e+00  8.493e-02 -25.843  < 2e-16 ***
## stateConnecticut           -2.191e+00  5.776e-02 -37.927  < 2e-16 ***
## stateDelaware              -2.799e+00  8.991e-02 -31.131  < 2e-16 ***
## stateDistrict of Columbia -2.891e+00  2.586e-01 -11.181  < 2e-16 ***
## stateFlorida                1.812e+00  3.945e-02  45.926  < 2e-16 ***
## stateGeorgia               -1.079e+00  1.119e-01  -9.646  < 2e-16 ***
## stateHawaii                -3.507e+00  3.543e-01  -9.898  < 2e-16 ***
## stateIdaho                 -3.638e+00  1.324e-01 -27.477  < 2e-16 ***
## stateIndiana               -1.928e+00  7.177e-02 -26.857  < 2e-16 ***
## stateIowa                  -3.168e+00  1.130e-01 -28.037  < 2e-16 ***
## stateKansas                -3.312e+00  9.023e-02 -36.701  < 2e-16 ***
## stateKentucky              -1.774e+00  9.241e-02 -19.201  < 2e-16 ***
## stateLouisiana             -2.806e+00  1.171e-01 -23.954  < 2e-16 ***
## stateMaine                 -2.838e+00  1.368e-01 -20.739  < 2e-16 ***
## stateMaryland              -1.259e+00  1.319e-01  -9.546  < 2e-16 ***
## stateMassachusetts         -1.156e+00  4.989e-02 -23.165  < 2e-16 ***
## stateMichigan              -6.912e-01  3.063e-02 -22.563  < 2e-16 ***
## stateMinnesota             -2.494e+00  7.099e-02 -35.139  < 2e-16 ***
## stateMississippi           -3.378e+00  1.468e-01 -23.005  < 2e-16 ***
## stateMissouri              -1.684e+00  5.868e-02 -28.706  < 2e-16 ***
## stateMontana               -3.679e+00  1.130e-01 -32.565  < 2e-16 ***
## stateNebraska              -4.140e+00  1.081e-01 -38.283  < 2e-16 ***
## stateNevada                -2.255e+00  5.875e-02 -38.387  < 2e-16 ***
## stateNew Hampshire         -2.517e+00  1.297e-01 -19.417  < 2e-16 ***
## stateNew Jersey            -1.165e+00  4.190e-02 -27.813  < 2e-16 ***
## stateNew Mexico            -2.422e+00  7.006e-02 -34.578  < 2e-16 ***
## stateNew York               1.779e+00  6.011e-02  29.598  < 2e-16 ***
## stateNorth Carolina        -6.578e-01  4.898e-02 -13.430  < 2e-16 ***
```

```
## stateNorth Dakota         -4.222e+00  1.179e-01 -35.818  < 2e-16 ***
## stateOhio                  2.152e-01  4.252e-02   5.061 4.17e-07 ***
## stateOklahoma             -2.096e+00  5.845e-02 -35.852  < 2e-16 ***
## stateOregon               -2.524e+00  9.416e-02 -26.807  < 2e-16 ***
## statePennsylvania         -8.648e-02  4.079e-02  -2.120 0.034005 *
## stateRhode Island         -2.579e+00  8.474e-02 -30.439  < 2e-16 ***
## stateSouth Carolina       -2.254e+00  8.467e-02 -26.623  < 2e-16 ***
## stateSouth Dakota         -4.036e+00  1.005e-01 -40.161  < 2e-16 ***
## stateTennessee            -1.439e+00  4.805e-02 -29.943  < 2e-16 ***
## stateTexas                 2.875e+00  7.769e-02  37.004  < 2e-16 ***
## stateUtah                 -2.233e+00  1.123e-01 -19.893  < 2e-16 ***
## stateVermont              -3.156e+00  1.402e-01 -22.512  < 2e-16 ***
## stateVirginia             -1.305e+00  5.776e-02 -22.585  < 2e-16 ***
## stateWashington           -1.561e+00  4.670e-02 -33.428  < 2e-16 ***
## stateWest Virginia        -1.901e+00  1.310e-01 -14.513  < 2e-16 ***
## stateWisconsin            -1.859e+00  8.038e-02 -23.129  < 2e-16 ***
## stateWyoming              -3.494e+00  1.335e-01 -26.179  < 2e-16 ***
## prop_white                 4.653e-01  7.031e-01   0.662 0.508153
## prescription_rate          4.351e-04  4.285e-04   1.016 0.309823
## unemp                      1.163e-02  3.027e-03   3.841 0.000122 ***
## total_pop                 -2.646e-07  5.384e-09 -49.148  < 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: 63933  on 509  degrees of freedom
## Residual deviance: 11075  on 446  degrees of freedom
## AIC: 15065
##
## Number of Fisher Scoring iterations: 4
```

```
## [1] 15064.91
```



The inclusion of 'expected_deaths' as an offset includes the `log()` of this variable in the model as a known constant. This incorporates this value into the estimation procedure, influencing the estimation of other coefficients (in this case 'inflate') in order to correct them to balance the model.

Besides, when comparing the model `fit1` with `fit2` we see that almost all coefficients associated to states are significant, different than the first model where some states doesn't appear to be, and the magnitide of coefficients differs from one model to the other.

(e) Investigate whether overdispersion is an issue in your current model.

In order to investigate the presence of overdispersion we will run a Chi-Square test in all three models adjusted so far, i.e., `fit1`, `fit1.1` and `fit2`.

```
##    model p-value
## 1   fit1       0
## 2 fit1.1       0
## 3   fit2       0
```

As all `p-value` are 'zero', we reject the hypotesis of **no overdispersion is present**.

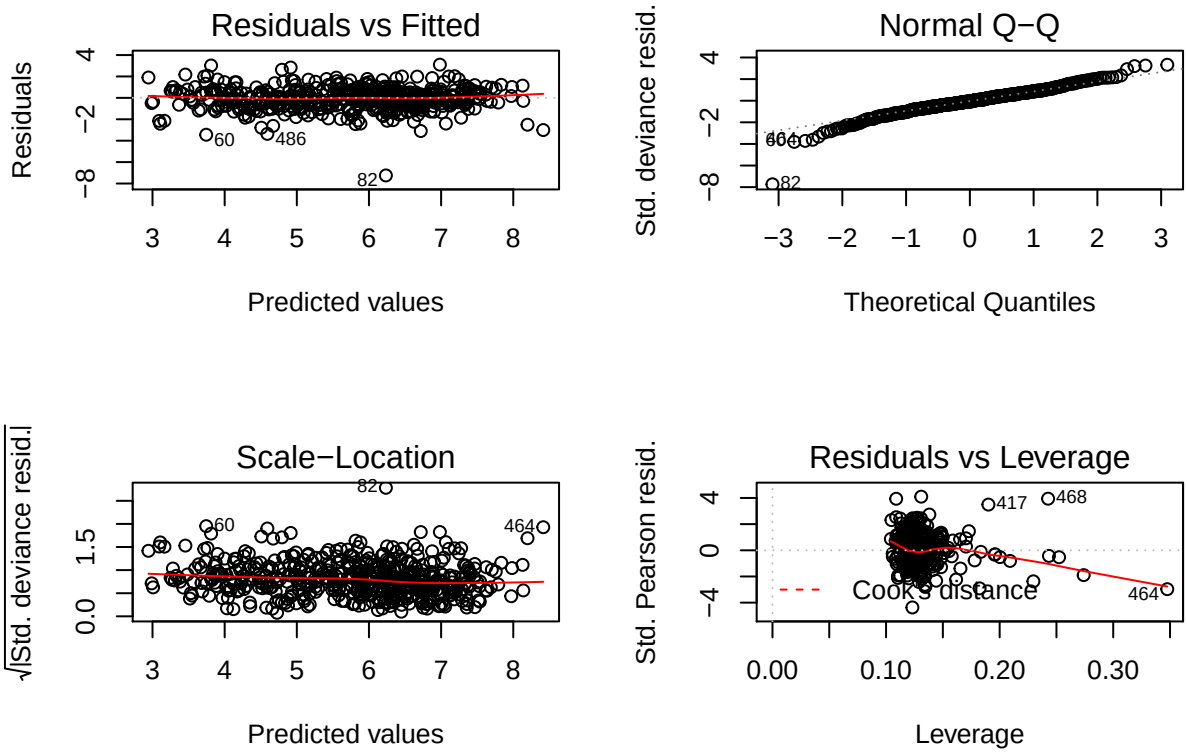
(f) If overdispersion is an issue, rerun your analysis using negative binomial regression (using the glm.nb function in the MASS library). Does this change the significance of your explanatory variables? Do a Likelihood Ratio Test to see which is the preferred model.

```
##
## Call:
## glm.nb(formula = deaths ~ year + state + prop_white + prescription_rate +
##     unemp + total_pop + expected_deaths, data = dt, init.theta = 21.90834277,
```

```
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.2452  -0.6054  -0.0620   0.5307   3.1011
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 3.892e+00  2.247e+00   1.732 0.083354 .
## year2009                    5.499e-02  6.557e-02   0.839 0.401670
## year2010                    1.129e-01  6.897e-02   1.637 0.101593
## year2011                    2.137e-01  6.350e-02   3.366 0.000763 ***
## year2012                    2.714e-01  5.825e-02   4.659 3.17e-06 ***
## year2013                    3.670e-01  5.590e-02   6.566 5.15e-11 ***
## year2014                    5.061e-01  5.481e-02   9.234  < 2e-16 ***
## year2015                    6.031e-01  5.897e-02  10.227  < 2e-16 ***
## year2016                    7.526e-01  6.680e-02  11.267  < 2e-16 ***
## year2017                    8.014e-01  8.032e-02   9.977  < 2e-16 ***
## stateAlabama               -3.526e+00  5.559e-01  -6.342 2.27e-10 ***
## stateAlaska                -6.426e+00  7.743e-01  -8.299  < 2e-16 ***
## stateArizona               -3.575e+00  4.239e-01  -8.432  < 2e-16 ***
## stateArkansas              -5.537e+00  6.119e-01  -9.049  < 2e-16 ***
## stateCalifornia             1.036e+01  1.537e+00   6.741 1.57e-11 ***
## stateColorado              -4.926e+00  5.477e-01  -8.993  < 2e-16 ***
## stateConnecticut           -4.963e+00  5.786e-01  -8.578  < 2e-16 ***
## stateDelaware              -6.019e+00  7.503e-01  -8.023 1.03e-15 ***
## stateDistrict of Columbia -4.056e+00  1.191e+00  -3.407 0.000657 ***
## stateFlorida                2.968e+00  4.276e-01   6.941 3.89e-12 ***
## stateGeorgia               -6.505e-02  4.574e-01  -0.142 0.886910
## stateHawaii                -2.407e+00  1.461e+00  -1.647 0.099491 .
## stateIdaho                 -8.469e+00  8.062e-01 -10.505  < 2e-16 ***
## stateIndiana               -4.010e+00  4.505e-01  -8.901  < 2e-16 ***
## stateIowa                  -7.087e+00  7.068e-01 -10.026  < 2e-16 ***
## stateKansas                -6.758e+00  6.589e-01 -10.256  < 2e-16 ***
## stateKentucky              -4.554e+00  5.931e-01  -7.679 1.61e-14 ***
## stateLouisiana             -3.176e+00  6.450e-01  -4.924 8.49e-07 ***
## stateMaine                 -8.060e+00  8.364e-01  -9.637  < 2e-16 ***
## stateMaryland              -1.199e+00  6.292e-01  -1.905 0.056780 .
## stateMassachusetts         -3.097e+00  4.132e-01  -7.496 6.58e-14 ***
## stateMichigan              -1.393e+00  2.167e-01  -6.430 1.28e-10 ***
## stateMinnesota             -5.060e+00  5.138e-01  -9.850  < 2e-16 ***
## stateMississippi           -3.924e+00  7.853e-01  -4.997 5.82e-07 ***
## stateMissouri              -3.653e+00  4.496e-01  -8.125 4.48e-16 ***
## stateMontana               -8.585e+00  7.886e-01 -10.886  < 2e-16 ***
## stateNebraska              -8.433e+00  7.388e-01 -11.414  < 2e-16 ***
## stateNevada                -4.672e+00  6.170e-01  -7.573 3.65e-14 ***
## stateNew Hampshire         -7.672e+00  8.228e-01  -9.325  < 2e-16 ***
## stateNew Jersey            -1.694e+00  2.833e-01  -5.980 2.24e-09 ***
## stateNew Mexico            -5.923e+00  6.724e-01  -8.808  < 2e-16 ***
## stateNew York               3.503e+00  4.538e-01   7.719 1.17e-14 ***
## stateNorth Carolina        -5.648e-01  2.688e-01  -2.101 0.035617 *
## stateNorth Dakota          -9.655e+00  8.032e-01 -12.021  < 2e-16 ***
## stateOhio                  -5.123e-01  1.941e-01  -2.640 0.008302 **
## stateOklahoma              -3.979e+00  5.581e-01  -7.130 1.00e-12 ***
```

```
## stateOregon               -5.586e+00  6.221e-01  -8.980  < 2e-16 ***
## statePennsylvania         -6.313e-01  1.750e-01  -3.608 0.000308 ***
## stateRhode Island         -7.113e+00  7.498e-01  -9.485  < 2e-16 ***
## stateSouth Carolina       -3.118e+00  5.633e-01  -5.535 3.11e-08 ***
## stateSouth Dakota         -8.976e+00  7.668e-01 -11.706  < 2e-16 ***
## stateTennessee            -2.656e+00  4.125e-01  -6.440 1.19e-10 ***
## stateTexas                 4.855e+00  8.313e-01   5.840 5.21e-09 ***
## stateUtah                 -6.148e+00  7.113e-01  -8.644  < 2e-16 ***
## stateVermont              -9.319e+00  8.786e-01 -10.607  < 2e-16 ***
## stateVirginia             -1.628e+00  3.471e-01  -4.690 2.73e-06 ***
## stateWashington           -3.172e+00  3.889e-01  -8.155 3.50e-16 ***
## stateWest Virginia        -6.319e+00  7.920e-01  -7.979 1.47e-15 ***
## stateWisconsin            -4.433e+00  5.139e-01  -8.625  < 2e-16 ***
## stateWyoming              -9.393e+00  8.511e-01 -11.037  < 2e-16 ***
## prop_white                 1.008e+01  2.639e+00   3.820 0.000133 ***
## prescription_rate         -5.002e-03  2.255e-03  -2.218 0.026534 *
## unemp                      5.407e-03  1.478e-02   0.366 0.714569
## total_pop                 -4.149e-07  6.598e-08  -6.288 3.21e-10 ***
## expected_deaths            4.128e-04  7.647e-05   5.398 6.74e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(21.9083) family taken to be 1)
##
##     Null deviance: 12091.93  on 509  degrees of freedom
## Residual deviance:   528.35  on 445  degrees of freedom
## AIC: 5939.3
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  21.91
##           Std. Err.:  1.54
##
##  2 x log-likelihood:  -5807.274
```

```
## [1] 5939.274
```



Performing the Likelihood Ratio test between the `fit3 - NegBin` and `fit2 - Poisson` shows that Negative Binomial model explains best the our data and is considered a better fitted model.

```
lrtest(fit2, fit3)
```

```
## Likelihood ratio test
##
## Model 1: deaths ~ year + state + prop_white + prescription_rate + unemp +
##     total_pop
## Model 2: deaths ~ year + state + prop_white + prescription_rate + unemp +
##     total_pop + expected_deaths
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  64 -7468.5
## 2  66 -2903.6  2 9129.6  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(g) Summarize your findings, giving the key insights into trends in opioid mortality over time and across states, and any factors that may be associated with these changes. What other variables may be of interest to investigate in future?

Summary of findings

From this analysis we can summarize the main findings as follows:

- The distribution of the variable of interest is adherent to a Poisson or Gamma distributions.
- For adjustments Poisson family or Negative Binomial are recommended in this type of study.

- **No. of Deaths** increases over the years;
- The main covariates with positive highest influence are **Expected No. of deaths**, **Total Population**
- **No. of Deaths** seems to decrease as **Prescription Rate** increases, in majority of states. Some exceptions were observed suggesting that other factors might be present for this reverse behaviour;
- **Proportion of White** has positive influence in **No. of Deaths** in increasing its rates, i.e., the higher prescription rate is present on the state, the less mortality is observed;
- **Unemployment Rate** has low influence on **No. of Deaths**
- Death-rate of **West Virginia** is the largest one among all US, followed by **Ohio**, **Maryland** and **New Hampshire** during 2008-2017 period.

For further investigation

Additional variables may be included in this study such as:

- **gender** to identify if the begaviour of deaths is different among the various group (including LGBT+ subgroups);
- Another aspect is **race** which can provide rich information of wether the etnic aspects presents potential influence of deaths by overdose from opioid abuse;
- The inclusin of **age** to identify the differences between each category and if its distribution may represent some insightful information;
- **socio-economic status** can provide other perspective of differences and influences on 'no. of deaths' distribution.
- Additional aspects may include **scholarship**, **family income** or - in a localized studies - **residential zone** might be of special interest for root-cause of opioid abuse and delineate prevention actions to reduce mortality rates in those areas.

This study can be also improved by performing cross-analysis with other researches to identify other demographics aspects (or even to define local actions) by inspecting other characteristics that might contribute to increasing the risk of opioid addiction.

For instance:

- Depression, anxiety or other mental issues;
- Family history of alchool abuse
- Other medical condition that leads to a long-term use of opioids (e.g., pain relief etc)