STA2201H Winter 2021 Assignment 1

Due: 5pm, 4 February 2021

What to hand in: .Rmd file and the compiled pdf

How to hand in: Submit files via Quercus

1 Overdispersion

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

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

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

Answer:

Since $Y|\theta \sim \text{Poisson}(\mu\theta)$. From the property of Poisson distribution, we could derive

$$E(Y|\theta) = \mu\theta, \quad Var(Y|\theta) = \mu\theta$$

Then by the law of total expectation:

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

By the law of total variance then:

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

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

Answer:

We assumed $\theta \sim Gamma(\alpha, \beta)$, $\quad Y|\theta \sim \text{Poisson}(\mu\theta)$. Then for the probability density function of y. We would have

$$f(y) = \int_0^\infty f(y|\theta)f(\theta)$$

$$\begin{aligned}
f(y) &= \int_0^\infty f(y|\theta)f(\theta) \\
&= \int_0^\infty \frac{(\mu\theta)^y e^{-\mu\theta}}{y!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} d\theta \\
&= \frac{\mu^y \beta^\alpha}{y!\Gamma(\alpha)} \cdot \int_0^\infty \theta^{y+\alpha-1} e^{-\theta(\mu+\beta)} d\theta \\
&= \frac{\mu^y \beta^\alpha}{y!\Gamma(\alpha)} \cdot \frac{\Gamma(y+\alpha)}{(\mu+\beta)^{y+\alpha}} \\
&= \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)} \cdot (\frac{\mu}{\mu+\beta})^y \cdot (\frac{\beta}{\mu+\beta})^\alpha \\
&= NB(\alpha, \frac{\mu}{\mu+\beta})
\end{aligned}$$

We could derive $f(y)$ to be the Negative Binomial PDF as above.

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

Answer: Given random variable $X \sim NB(r, p)$. We could derive

$$E(X) = \frac{pr}{1-p}$$

and

$$Var(X) = \frac{pr}{(1-p)^2}$$

. Given the distribution we have above, we than could derive: $\frac{\mu\alpha}{\beta} = \mu$ and $\frac{\alpha\mu^2 + \alpha\beta\mu}{\beta^2} = \mu(1 + \mu\sigma^2)$. Solve two equations we could have $\alpha = \beta = \frac{1}{\sigma^2}$.

2 Opioid mortality in the US

The following questions relate to the `opioids` dataset, which you can find in the `data` folder of the 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`.

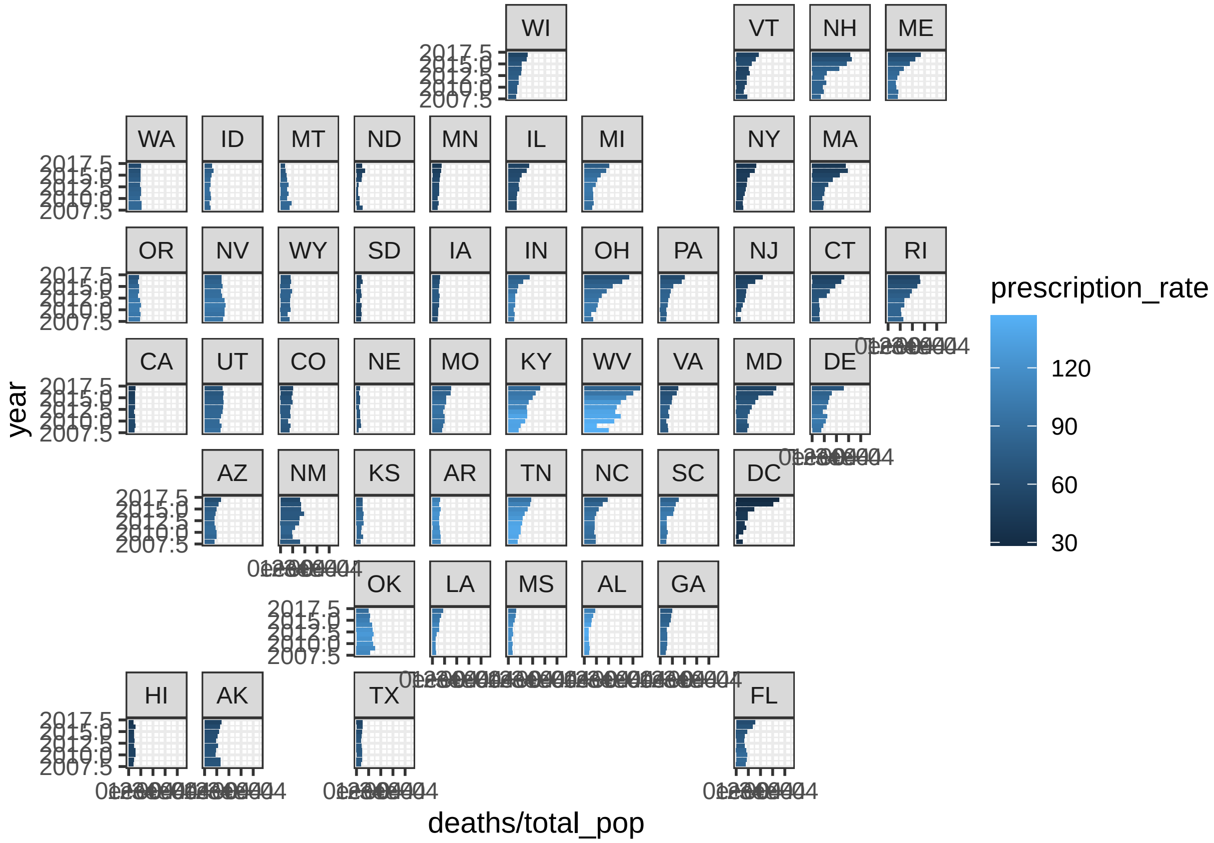
Please make sure to clearly explain any findings or observations you make, rather than just handing in code and output. You will be assessed not only on the code but also on how you communicate your findings with a combination of writing and analysis.

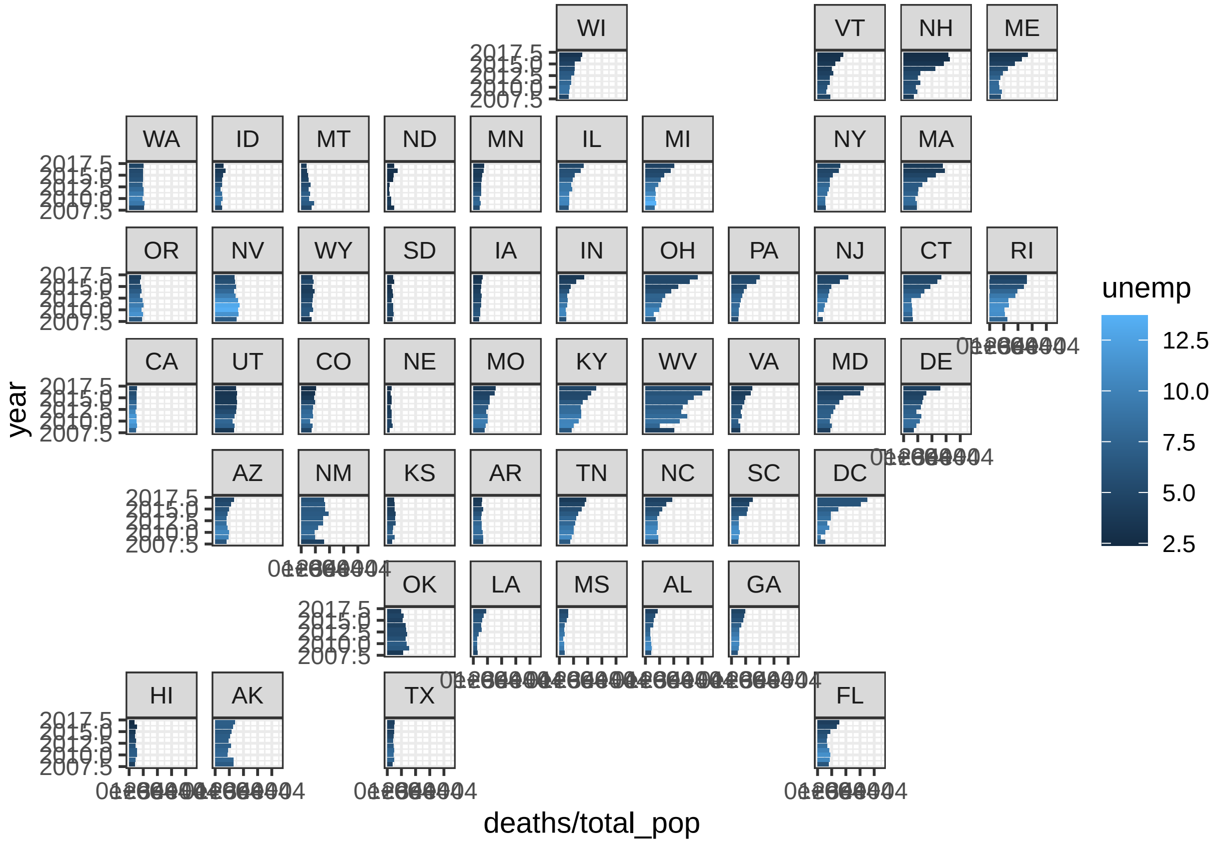
a) Perform some exploratory data analysis (EDA) using this dataset, and briefly summarize in words, tables and charts your main observations. You may use whatever tools or packages you wish. You may want to explore the `geofacet` package, which plots US state facets in the correct geographic orientation.

Answer:

We found some patterns between the death rate and the other factors. Since different states could have different populations. So we use death_rate = deaths/population to describe the loss in population. Then we could find

- The west coast often have lower death rate than the east coast and the middle area.
- Not only that, the middle area and the east coast death rates keeps increasing.
- The death rate do not have a strong correlation of the white people proportion. More likely, they have some relationships with the unemployment rates, in NV for example.
- In the middle we observe a high prescription rate but the it decreases as time goes. However, before 2015 there are some positive relationship between the presciption rate and the death rates.


WI
VT
NH
ME
WA
ID
MT
ND
MN
IL
MI
NY
MA
OR
NV
WY
SD
IA
IN
OH
PA
NJ
CT
RI
CA
UT
CO
NE
MO
KY
WV
VA
MD
DE
AZ
NM
KS
AR
TN
NC
SC
DC
OK
LA
MS
AL
GA
HI
AK
TX
FL
2017.5
2015.0
2012.5
2010.0
2007.5
year
deaths/total_pop
prescription_rate
120
90
60
30


WI
VT
NH
ME
WA
ID
MT
ND
MN
IL
MI
NY
MA
OR
NV
WY
SD
IA
IN
OH
PA
NJ
CT
RI
CA
UT
CO
NE
MO
KY
WV
VA
MD
DE
AZ
NM
KS
AR
TN
NC
SC
DC
OK
LA
MS
AL
GA
HI
AK
TX
FL
2017.5
2015.0
2012.5
2010.0
2007.5
year
deaths/total_pop
unemp
12.5
10.0
7.5
5.0
2.5

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. 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.

Answer:

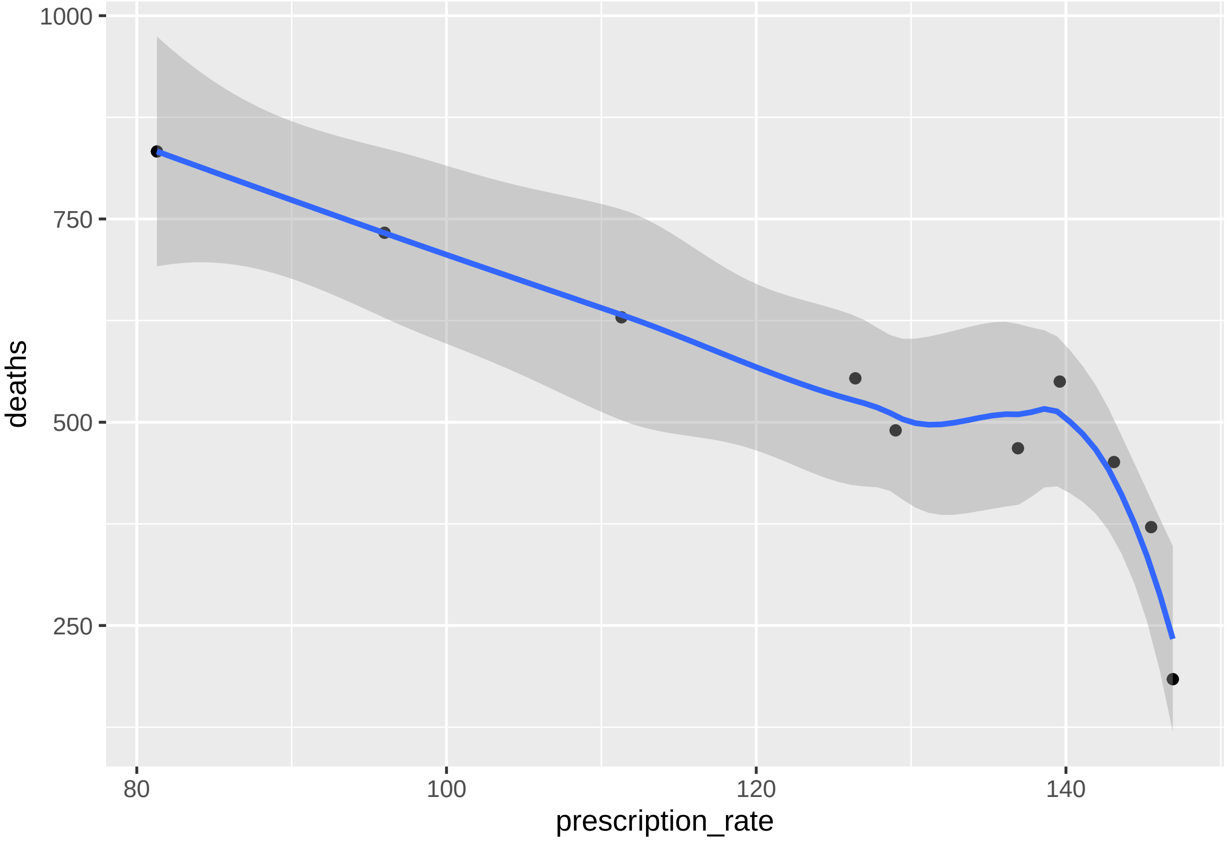
- We first include the log(total_pop) as our offset variable and all the other potential relevant variables. Then we use StepAIC function to select the optimum variables. *The kept variables are unemp, prescription_rate, year and state.*
- The kept variables fit our initial expection in part (a). Moreover, by analyzing the coefficients. *We could find the state with the highest opioid mortality will be West Virginia (1.657).* Combine with the visualization in part(a), the prescription rate plays an important role in the death rate in West Virginia. The visualization of West Virginia specifically proved this founding.

```
## Start:  AIC=18469.93
## deaths ~ offset(log(total_pop)) + year + state + unemp + prescription_rate +
##     prop_white + expected_deaths
##
##                     Df Deviance   AIC
## - prop_white         1    14497 18469
## <none>                    14496 18470
## - unemp              1    14547 18519
## - expected_deaths    1    14636 18608
## - prescription_rate  1    15404 19376
## - year               1    15981 19953
## - state             50    52043 55917
##
## Step:  AIC=18469.31
## deaths ~ year + state + unemp + prescription_rate + expected_deaths +
##     offset(log(total_pop))
##
##                     Df Deviance   AIC
## <none>                    14497 18469
## - unemp              1    14547 18517
## - expected_deaths    1    14637 18607
## - prescription_rate  1    15404 19374
## - year               1    18310 22280
## - state             50    53153 57025
```

```
##
## Call:
## glm(formula = deaths ~ offset(log(total_pop)) + year + state +
##     unemp + prescription_rate, family = poisson, data = opioids)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -26.674   -2.842   -0.442    2.198   21.783
```

```
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -1.477e+02  2.335e+00 -63.248  < 2e-16 ***
## year                       6.911e-02  1.153e-03  59.963  < 2e-16 ***
## stateAlabama               5.433e-03  3.150e-02   0.172  0.86307
## stateAlaska                1.631e-01  3.620e-02   4.505 6.63e-06 ***
## stateArizona               1.427e-01  1.654e-02   8.628  < 2e-16 ***
## stateArkansas              9.030e-02  3.153e-02   2.864  0.00418 **
## stateCalifornia           -7.342e-01  1.196e-02 -61.375  < 2e-16 ***
## stateColorado             -9.970e-02  1.801e-02  -5.536 3.10e-08 ***
## stateConnecticut           2.871e-01  1.743e-02  16.474  < 2e-16 ***
## stateDelaware              5.662e-01  3.165e-02  17.889  < 2e-16 ***
## stateDistrict of Columbia  6.151e-02  3.638e-02   1.691  0.09084 .
## stateFlorida               7.316e-02  1.256e-02   5.825 5.73e-09 ***
## stateGeorgia              -1.758e-01  1.701e-02 -10.335  < 2e-16 ***
## stateHawaii               -9.079e-01  4.078e-02 -22.267  < 2e-16 ***
## stateIdaho                -4.406e-01  3.725e-02 -11.829  < 2e-16 ***
## stateIndiana               9.560e-02  2.063e-02   4.633 3.60e-06 ***
## stateIowa                 -5.642e-01  2.675e-02 -21.092  < 2e-16 ***
## stateKansas               -4.845e-01  2.959e-02 -16.373  < 2e-16 ***
## stateKentucky              1.036e+00  2.271e-02  45.613  < 2e-16 ***
## stateLouisiana            -2.547e-01  2.763e-02  -9.218  < 2e-16 ***
## stateMaine                 4.439e-01  2.700e-02  16.440  < 2e-16 ***
## stateMaryland              5.234e-01  1.411e-02  37.080  < 2e-16 ***
## stateMassachusetts         4.516e-01  1.345e-02  33.566  < 2e-16 ***
## stateMichigan              3.990e-01  1.648e-02  24.208  < 2e-16 ***
## stateMinnesota            -6.131e-01  2.044e-02 -29.988  < 2e-16 ***
## stateMississippi          -3.514e-01  3.429e-02 -10.246  < 2e-16 ***
## stateMissouri              3.693e-01  1.783e-02  20.715  < 2e-16 ***
## stateMontana              -3.781e-01  4.335e-02  -8.721  < 2e-16 ***
## stateNebraska             -1.225e+00  4.556e-02 -26.885  < 2e-16 ***
## stateNevada                7.149e-01  1.995e-02  35.842  < 2e-16 ***
## stateNew Hampshire         7.052e-01  2.393e-02  29.476  < 2e-16 ***
## stateNew Jersey           -1.913e-01  1.470e-02 -13.013  < 2e-16 ***
## stateNew Mexico            5.089e-01  2.064e-02  24.652  < 2e-16 ***
## stateNew York             -2.070e-01  1.242e-02 -16.666  < 2e-16 ***
## stateNorth Carolina        3.746e-01  1.618e-02  23.159  < 2e-16 ***
## stateNorth Dakota         -1.099e+00  6.407e-02 -17.159  < 2e-16 ***
## stateOhio                  8.165e-01  1.468e-02  55.600  < 2e-16 ***
## stateOklahoma              7.183e-01  2.447e-02  29.352  < 2e-16 ***
## stateOregon                1.577e-01  2.142e-02   7.362 1.81e-13 ***
## statePennsylvania          7.986e-02  1.406e-02   5.679 1.36e-08 ***
## stateRhode Island          7.019e-01  2.518e-02  27.876  < 2e-16 ***
## stateSouth Carolina        1.627e-01  2.154e-02   7.552 4.30e-14 ***
## stateSouth Dakota         -9.801e-01  5.630e-02 -17.409  < 2e-16 ***
## stateTennessee             8.472e-01  2.461e-02  34.431  < 2e-16 ***
## stateTexas                -7.407e-01  1.335e-02 -55.482  < 2e-16 ***
## stateUtah                   5.367e-01  1.988e-02  26.990  < 2e-16 ***
## stateVermont                3.366e-02  3.973e-02   0.847  0.39685    
## stateVirginia              -4.514e-02  1.589e-02  -2.840  0.00451 ** 
## stateWashington             1.360e-01  1.580e-02   8.613  < 2e-16 ***
## stateWest Virginia          1.657e+00  2.519e-02  65.758  < 2e-16 ***
## stateWisconsin              1.067e-01  1.652e-02   6.460 1.05e-10 ***
## stateWyoming               -6.693e-02  4.850e-02  -1.380  0.16754    
## unemp                      -6.480e-03  1.572e-03  -4.121 3.77e-05 ***
## prescription_rate          -9.881e-03  3.239e-04 -30.507  < 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: 14637  on 456  degrees of freedom
## AIC: 18607
## 
## Number of Fisher Scoring iterations: 4
```

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



c) What's an issue with using population as an offset, given the limited information available in this dataset?

Answer:

When use population as an offset. We potentially want to discover the relationships between deaths/total_pop and the other factors. Just as we have found before, there is a strong relation between such rate and the geographical differences (West coast and east coast). Because of that, it is more difficult to find the relation between death and the prescription_rates. Which are our primary interests.

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

Answer:

The weights on different states decrease and the weight on prescription_rate increases as our expected. Since deaths/expected_deaths are less more likely to be affect by the geographical differences and its following impacts. Medical services, economic reasons and other factors for example.

```
##
## Call:
## glm(formula = deaths ~ offset(log(expected_deaths)) + year +
##     state + unemp + prescription_rate, family = poisson, data = opioids)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -27.2381   -2.5419    0.0689    2.3719   19.3086
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -4.061315   2.328101  -1.744 0.081076 .
## year                       0.002067   0.001149   1.799 0.071964 .
## stateAlabama              -0.514028   0.031480 -16.329  < 2e-16 ***
## stateAlaska                0.116496   0.036193   3.219 0.001287 **
## stateArizona               0.051033   0.016522   3.089 0.002009 **
## stateArkansas             -0.292046   0.031536  -9.261  < 2e-16 ***
## stateCalifornia           -0.681316   0.011970 -56.919  < 2e-16 ***
## stateColorado             -0.136833   0.017993  -7.605 2.86e-14 ***
## stateConnecticut           0.282915   0.017422  16.239  < 2e-16 ***
## stateDelaware              0.389289   0.031629  12.308  < 2e-16 ***
## stateDistrict of Columbia  0.123587   0.036365   3.399 0.000678 ***
## stateFlorida               0.002930   0.012555   0.233 0.815475
## stateGeorgia              -0.358595   0.016996 -21.099  < 2e-16 ***
## stateHawaii               -0.735018   0.040775 -18.026  < 2e-16 ***
## stateIdaho                -0.537436   0.037223 -14.438  < 2e-16 ***
## stateIndiana              -0.145952   0.020608  -7.082 1.42e-12 ***
## stateIowa                 -0.507502   0.026731 -18.985  < 2e-16 ***
## stateKansas               -0.577737   0.029554 -19.549  < 2e-16 ***
## stateKentucky              0.636010   0.022905  27.767  < 2e-16 ***
## stateLouisiana            -0.588779   0.027582 -21.346  < 2e-16 ***
## stateMaine                 0.334842   0.026988  12.407  < 2e-16 ***
## stateMaryland              0.501884   0.014093  35.613  < 2e-16 ***
## stateMassachusetts         0.485178   0.013449  36.076  < 2e-16 ***
## stateMichigan              0.164968   0.016438  10.036  < 2e-16 ***
## stateMinnesota            -0.509449   0.020434 -24.932  < 2e-16 ***
## stateMississippi          -0.731809   0.034275 -21.351  < 2e-16 ***
## stateMissouri              0.203923   0.017782  11.468  < 2e-16 ***
## stateMontana              -0.448327   0.043332 -10.346  < 2e-16 ***
## stateNebraska             -1.164589   0.045533 -25.577  < 2e-16 ***
## stateNevada                0.459476   0.019935  23.048  < 2e-16 ***
## stateNew Hampshire         0.640287   0.023891  26.800  < 2e-16 ***
## stateNew Jersey           -0.154377   0.014698 -10.503  < 2e-16 ***
## stateNew Mexico            0.477914   0.020629  23.167  < 2e-16 ***
## stateNew York             -0.092330   0.012420  -7.434 1.05e-13 ***
## stateNorth Carolina        0.160662   0.016142   9.953  < 2e-16 ***
## stateNorth Dakota         -0.967678   0.064053 -15.107  < 2e-16 ***
## stateOhio                  0.635529   0.014662  43.346  < 2e-16 ***
## stateOklahoma              0.401012   0.024399  16.436  < 2e-16 ***
## stateOregon               -0.056616   0.021405  -2.645 0.008170 ** 
## statePennsylvania         -0.004617   0.014033  -0.329 0.742164    
## stateRhode Island          0.601465   0.025178  23.889  < 2e-16 ***
## stateSouth Carolina       -0.091477   0.021515  -4.252 2.12e-05 ***
## stateSouth Dakota         -0.828887   0.056282 -14.727  < 2e-16 ***
## stateTennessee             0.384963   0.024671  15.604  < 2e-16 ***
## stateTexas                -0.741250   0.013321 -55.645  < 2e-16 ***
## stateUtah                  0.503522   0.019825  25.398  < 2e-16 ***
## stateVermont               0.147470   0.039723   3.712 0.000205 ***
## stateVirginia             -0.096898   0.015853  -6.112 9.83e-10 ***
## stateWashington            0.028138   0.015779   1.783 0.074551 .  
## stateWest Virginia         1.225700   0.025567  47.941  < 2e-16 ***
## stateWisconsin             0.087526   0.016497   5.306 1.12e-07 ***
## stateWyoming              -0.133888   0.048471  -2.762 0.005741 ** 
## unemp                      0.012056   0.001553   7.764 8.22e-15 ***
## prescription_rate         -0.001928   0.000324  -5.951 2.67e-09 ***
## ---
## 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: 13612  on 456  degrees of freedom
## AIC: 17582
## 
## Number of Fisher Scoring iterations: 4
```

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

Answer:

From the residual deviance we could see it is 13612 with 456 degrees of freedom. Consider the poisson family should be 1. However, the ratio instead is 29.85 >> 1. We use dispersion tests to verify our foundings. Since the p-value is very small. We reject the null hypothesis and prefer the alternative hypothesis. The true alpha is greater than 0 and indicates there is an over dispersion.

```
##
##  Overdispersion test
##
## data:  poisson_model
## z = 10.428, p-value < 2.2e-16
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##    alpha
## 24.74271
```

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

Answer:

After we use the negative binomial regression to fit the data. We could observe much less variables are statistically significant. Which make sense since over dispersion falsify some parameters to have samller values. We again use a likelihood ratio test to verify our findings. From the results we could conclude negative binomial regression has higher log likelihood than poisson model.

```
##
## Call:
## glm.nb(formula = deaths ~ offset(log(expected_deaths)) + year +
##     state + unemp + prescription_rate, data = opioids, init.theta = 18.70368876,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7642  -0.6107  -0.0264   0.5475   2.9082
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              20.4447263 11.1967852   1.826 0.067858 .
## year                     -0.0100999  0.0055258  -1.828 0.067584 .
## stateAlabama             -0.5862558  0.1675259  -3.499 0.000466 ***
## stateAlaska               0.1714686  0.1100821   1.558 0.119318
## stateArizona              0.0848036  0.1105451   0.767 0.442998
## stateArkansas            -0.2923561  0.1526525  -1.915 0.055470 .
## stateCalifornia          -0.5780032  0.1059707  -5.454 4.91e-08 ***
## stateColorado            -0.1099236  0.1077948  -1.020 0.307848
## stateConnecticut          0.2111777  0.1056916   1.998 0.045711 *
## stateDelaware             0.3429641  0.1241078   2.763 0.005720 **
## stateDistrict of Columbia  0.0449588  0.1195055   0.376 0.706763
## stateFlorida              0.0102611  0.1081241   0.095 0.924393
## stateGeorgia             -0.3662369  0.1133159  -3.232 0.001229 **
## stateHawaii               -0.6791814  0.1145772  -5.928 3.07e-09 ***
## stateIdaho                -0.5407280  0.1216480  -4.445 8.79e-06 ***
## stateIndiana              -0.2310208  0.1274536  -1.813 0.069895 .
## stateIowa                 -0.4964792  0.1114059  -4.456 8.33e-06 ***
## stateKansas               -0.5704575  0.1202135  -4.745 2.08e-06 ***
## stateKentucky              0.5836501  0.1499141   3.893 9.89e-05 ***
## stateLouisiana            -0.6690568  0.1398878  -4.783 1.73e-06 ***
## stateMaine                 0.2507413  0.1168509   2.146 0.031887 *
## stateMaryland              0.4308936  0.1066705   4.039 5.36e-05 ***
## stateMassachusetts         0.4318382  0.1053733   4.098 4.16e-05 ***
## stateMichigan              0.1140340  0.1186350   0.961 0.336443
## stateMinnesota            -0.4919327  0.1073762  -4.581 4.62e-06 ***
## stateMississippi          -0.7683676  0.1450836  -5.296 1.18e-07 ***
## stateMissouri              0.1912094  0.1177437   1.624 0.104387
## stateMontana              -0.3865561  0.1212355  -3.188 0.001430 **
## stateNebraska             -1.1529678  0.1201978  -9.592  < 2e-16 ***
## stateNevada                0.5480909  0.1172280   4.675 2.93e-06 ***
## stateNew Hampshire         0.5390615  0.1163777   4.632 3.62e-06 ***
## stateNew Jersey           -0.2386928  0.1047984  -2.278 0.022748 *
## stateNew Mexico            0.5088683  0.1086471   4.684 2.82e-06 ***
## stateNew York             -0.1009342  0.1068061  -0.945 0.344647
## stateNorth Carolina        0.1402803  0.1174491   1.194 0.232325
## stateNorth Dakota         -1.0098233  0.1280506  -7.886 3.12e-15 ***
## stateOhio                  0.5370377  0.1188369   4.519 6.21e-06 ***
## stateOklahoma              0.4213369  0.1491608   2.825 0.004732 **
## stateOregon                0.0116771  0.1196698   0.098 0.922268
## statePennsylvania         -0.0832726  0.1098353  -0.758 0.448356
## stateRhode Island          0.6005944  0.1090373   5.508 3.63e-08 ***
## stateSouth Carolina       -0.1566708  0.1235300  -1.268 0.204697
## stateSouth Dakota         -0.8155744  0.1219200  -6.689 2.24e-11 ***
## stateTennessee             0.3221588  0.1599831   2.014 0.044040 *
## stateTexas                -0.6888530  0.1067573  -6.453 1.10e-10 ***
## stateUtah                  0.5319959  0.1169291   4.550 5.37e-06 ***
## stateVermont               0.1182165  0.1140155   1.037 0.299808
## stateVirginia             -0.1394369  0.1099946  -1.268 0.204916
## stateWashington            0.1000321  0.1090860   0.917 0.359141
## stateWest Virginia         1.1604992  0.1620341   7.162 7.95e-13 ***
## stateWisconsin             0.0662997  0.1079673   0.614 0.539168
## stateWyoming              -0.1287216  0.1231804  -1.045 0.296030
## unemp                     -0.0008672  0.0086109  -0.101 0.919783
## prescription_rate         -0.0007974  0.0018500  -0.431 0.666437
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(18.7037) family taken to be 1)
##
##     Null deviance: 2387.35  on 509  degrees of freedom
## Residual deviance:  523.71  on 456  degrees of freedom
## AIC: 5985.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  18.70
##           Std. Err.:  1.29
##
##  2 x log-likelihood:  -5875.766
```

```
## Likelihood ratio test
##
## Model 1: deaths ~ offset(log(expected_deaths)) + year + state + unemp +
##     prescription_rate
## Model 2: deaths ~ offset(log(expected_deaths)) + year + state + unemp +
##     prescription_rate
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1  54 -8736.9
## 2  55 -2937.9  1 11598  < 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?

Answer:

- From the prelimary investigation on the data. We raised couple hypothesis, the death rate are affect by geometrical differences, year, prescription_rate and unemployment rate. For the variables selection part we then verify those findings. The west coast has lower death rate across the country. Moreover, the death rate increase as time goes.
- Based on those findings. We use a negative binomial regression model with expected death as offset variable to aovid overdispersion problem and the affect by total population in states. Further more, from the coefficients we found the prescription rate has a negative effect on the populations. The unemployment rate and the time has a positive effect on the death rate instead. The differences between states are even larger. For example, West Virginia has the highest mortality rate across the country based on the coefficients. The Nebraska has the lowest.
- We could then raise some other factors might be interesting. We want to ask why the state differences could affect the mortality rate that much. To answer this question, we want to investigate the electircal health record of the patients who are prescribed across different states. Which contain more information of individual differences.

3 Gompertz

Gompertz hazards are of the form

$$\lambda(t) = \alpha e^{\beta t}$$

for $t \in [0, \infty)$ with $\alpha, \beta > 0$. It is named after Benjamin Gompertz, who suggested a similar form to capture a 'law of human mortality' in 1825.

This question uses the `ON_mortality.RDS` file in the `data` folder of the class repo. This file contains hazard rates (`hx`) and density of deaths (`dx`) by age and year for Ontario. Note that in this case, the survival times we are interested in are age.

```
## Warning: NAs introduced by coercion
```

a) Find an expression in terms of α and β for the modal age at death.

Answer:

To find the mode age of death. We need to find the maximum density of the potential distribution. To do that, we need to find $f(t)$ first.

$$\begin{aligned} f(t) &= \lambda(t)s(t) \\ &= \alpha e^{\beta t} \cdot exp\{-\int_0^t \lambda(x)dx\} \\ &= \alpha e^{\beta t} \cdot exp\{-\int_0^t \alpha e^{\beta x}dx\} \\ &= \alpha e^{\beta t} \cdot exp\{-(\frac{\alpha}{\beta}e^{\beta x}|_0^t)\} \\ &= \alpha e^{\beta t} \cdot exp\{-\frac{\alpha}{\beta}(e^{\beta t} - 1)\} \end{aligned}$$

Then the derivative of $log(f(x))$ could be dervived

$$\begin{aligned} \frac{dlog(f(t))}{dt} &= \frac{d}{dt}(log(\alpha) + \beta t - \frac{\alpha}{\beta}(e^{\beta t} - 1)) \\ &= \beta - \alpha e^{\beta t} \\ &= 0 \end{aligned}$$

Then we could have optimum $\hat{t} = \frac{log(\beta) - log(\alpha)}{\beta}$

b) For every year, estimate α, β and the mode age at death.

Answer

The estimated Value are given below:

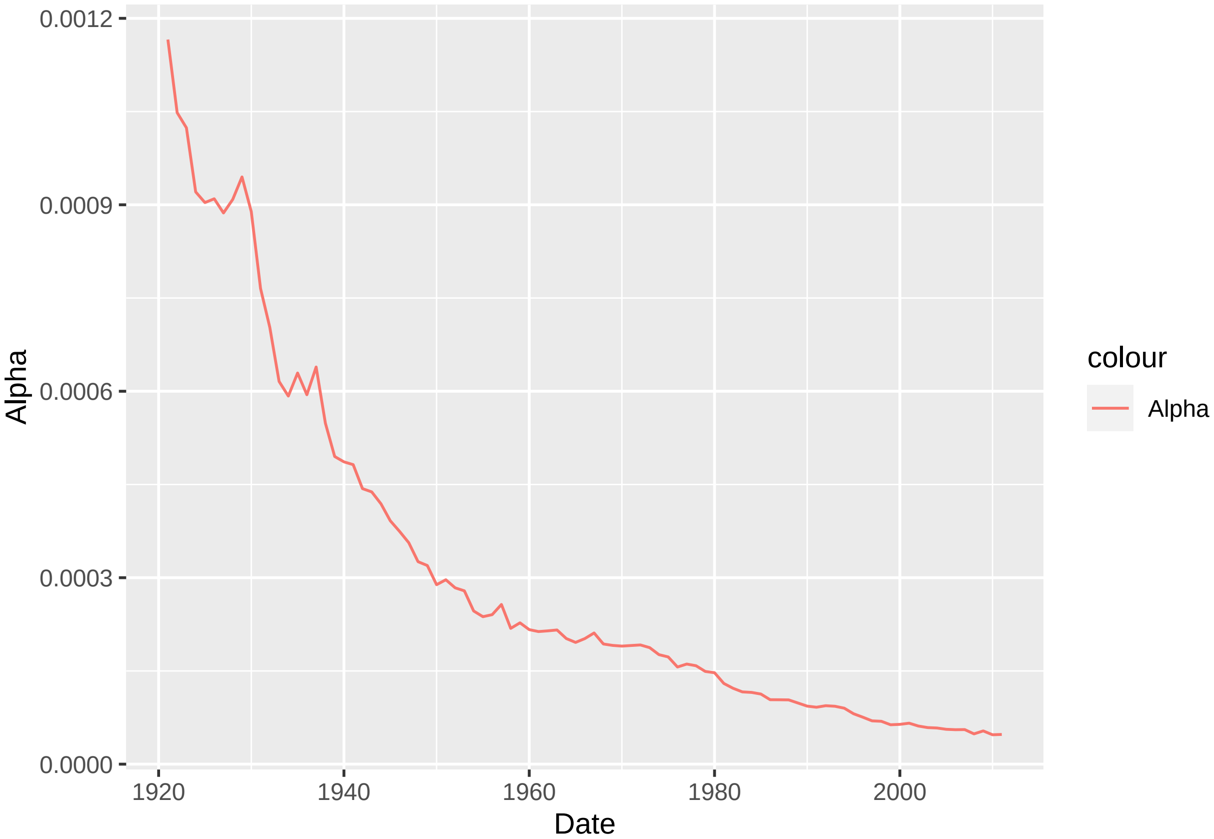
```
## [1] "In 1921 , The estimated alpha is ~0.00117,beta is 0.055 and mode age is 69.986"
## [1] "In 1922 , The estimated alpha is ~0.00105,beta is 0.056 and mode age is 70.581"
## [1] "In 1923 , The estimated alpha is ~0.00102,beta is 0.057 and mode age is 70.211"
## [1] "In 1924 , The estimated alpha is ~0.00092,beta is 0.058 and mode age is 71.444"
## [1] "In 1925 , The estimated alpha is ~9e-04,beta is 0.058 and mode age is 71.471"
## [1] "In 1926 , The estimated alpha is ~0.00091,beta is 0.059 and mode age is 70.934"
## [1] "In 1927 , The estimated alpha is ~0.00089,beta is 0.059 and mode age is 71.293"
## [1] "In 1928 , The estimated alpha is ~0.00091,beta is 0.059 and mode age is 70.311"
## [1] "In 1929 , The estimated alpha is ~0.00094,beta is 0.058 and mode age is 70.552"
## [1] "In 1930 , The estimated alpha is ~0.00089,beta is 0.059 and mode age is 71.08"
## [1] "In 1931 , The estimated alpha is ~0.00077,beta is 0.06 and mode age is 72.416"
## [1] "In 1932 , The estimated alpha is ~7e-04,beta is 0.062 and mode age is 72.374"
## [1] "In 1933 , The estimated alpha is ~0.00062,beta is 0.063 and mode age is 73.389"
## [1] "In 1934 , The estimated alpha is ~0.00059,beta is 0.063 and mode age is 73.786"
## [1] "In 1935 , The estimated alpha is ~0.00063,beta is 0.063 and mode age is 73.535"
## [1] "In 1936 , The estimated alpha is ~0.00059,beta is 0.064 and mode age is 73.282"
## [1] "In 1937 , The estimated alpha is ~0.00064,beta is 0.063 and mode age is 73.092"
## [1] "In 1938 , The estimated alpha is ~0.00055,beta is 0.064 and mode age is 74.216"
## [1] "In 1939 , The estimated alpha is ~0.00049,beta is 0.066 and mode age is 74.491"
## [1] "In 1940 , The estimated alpha is ~0.00049,beta is 0.066 and mode age is 74.444"
## [1] "In 1941 , The estimated alpha is ~0.00048,beta is 0.066 and mode age is 74.517"
## [1] "In 1942 , The estimated alpha is ~0.00044,beta is 0.067 and mode age is 75.167"
## [1] "In 1943 , The estimated alpha is ~0.00044,beta is 0.068 and mode age is 74.579"
## [1] "In 1944 , The estimated alpha is ~0.00042,beta is 0.067 and mode age is 75.666"
## [1] "In 1945 , The estimated alpha is ~0.00039,beta is 0.068 and mode age is 76.12"
## [1] "In 1946 , The estimated alpha is ~0.00037,beta is 0.068 and mode age is 76.413"
## [1] "In 1947 , The estimated alpha is ~0.00036,beta is 0.069 and mode age is 76.404"
## [1] "In 1948 , The estimated alpha is ~0.00033,beta is 0.07 and mode age is 76.582"
## [1] "In 1949 , The estimated alpha is ~0.00032,beta is 0.07 and mode age is 76.723"
## [1] "In 1950 , The estimated alpha is ~0.00029,beta is 0.072 and mode age is 77.047"
## [1] "In 1951 , The estimated alpha is ~3e-04,beta is 0.071 and mode age is 77.261"
## [1] "In 1952 , The estimated alpha is ~0.00028,beta is 0.071 and mode age is 77.792"
## [1] "In 1953 , The estimated alpha is ~0.00028,beta is 0.071 and mode age is 77.587"
## [1] "In 1954 , The estimated alpha is ~0.00025,beta is 0.073 and mode age is 78.326"
## [1] "In 1955 , The estimated alpha is ~0.00024,beta is 0.073 and mode age is 78.463"
## [1] "In 1956 , The estimated alpha is ~0.00024,beta is 0.073 and mode age is 78.423"
## [1] "In 1957 , The estimated alpha is ~0.00026,beta is 0.072 and mode age is 78.277"
## [1] "In 1958 , The estimated alpha is ~0.00022,beta is 0.074 and mode age is 78.679"
## [1] "In 1959 , The estimated alpha is ~0.00023,beta is 0.074 and mode age is 78.53"
## [1] "In 1960 , The estimated alpha is ~0.00022,beta is 0.074 and mode age is 78.749"
## [1] "In 1961 , The estimated alpha is ~0.00021,beta is 0.074 and mode age is 79.052"
## [1] "In 1962 , The estimated alpha is ~0.00021,beta is 0.074 and mode age is 79.03"
## [1] "In 1963 , The estimated alpha is ~0.00022,beta is 0.074 and mode age is 78.76"
## [1] "In 1964 , The estimated alpha is ~2e-04,beta is 0.074 and mode age is 79.59"
## [1] "In 1965 , The estimated alpha is ~2e-04,beta is 0.075 and mode age is 79.122"
## [1] "In 1966 , The estimated alpha is ~2e-04,beta is 0.074 and mode age is 79.437"
## [1] "In 1967 , The estimated alpha is ~0.00021,beta is 0.074 and mode age is 79.518"
## [1] "In 1968 , The estimated alpha is ~0.00019,beta is 0.075 and mode age is 79.647"
## [1] "In 1969 , The estimated alpha is ~0.00019,beta is 0.075 and mode age is 79.906"
## [1] "In 1970 , The estimated alpha is ~0.00019,beta is 0.074 and mode age is 80.308"
## [1] "In 1971 , The estimated alpha is ~0.00019,beta is 0.074 and mode age is 80.337"
## [1] "In 1972 , The estimated alpha is ~0.00019,beta is 0.074 and mode age is 80.117"
## [1] "In 1973 , The estimated alpha is ~0.00019,beta is 0.075 and mode age is 80.118"
## [1] "In 1974 , The estimated alpha is ~0.00018,beta is 0.075 and mode age is 80.371"
## [1] "In 1975 , The estimated alpha is ~0.00017,beta is 0.075 and mode age is 80.728"
## [1] "In 1976 , The estimated alpha is ~0.00016,beta is 0.076 and mode age is 81.198"
## [1] "In 1977 , The estimated alpha is ~0.00016,beta is 0.076 and mode age is 81.36"
## [1] "In 1978 , The estimated alpha is ~0.00016,beta is 0.075 and mode age is 81.829"
## [1] "In 1979 , The estimated alpha is ~0.00015,beta is 0.076 and mode age is 82.154"
## [1] "In 1980 , The estimated alpha is ~0.00015,beta is 0.076 and mode age is 82.148"
## [1] "In 1981 , The estimated alpha is ~0.00013,beta is 0.077 and mode age is 82.685"
## [1] "In 1982 , The estimated alpha is ~0.00012,beta is 0.078 and mode age is 82.89"
## [1] "In 1983 , The estimated alpha is ~0.00012,beta is 0.078 and mode age is 83.131"
## [1] "In 1984 , The estimated alpha is ~0.00012,beta is 0.078 and mode age is 83.437"
## [1] "In 1985 , The estimated alpha is ~0.00011,beta is 0.079 and mode age is 83.215"
## [1] "In 1986 , The estimated alpha is ~1e-04,beta is 0.08 and mode age is 83.433"
## [1] "In 1987 , The estimated alpha is ~1e-04,beta is 0.079 and mode age is 83.643"
## [1] "In 1988 , The estimated alpha is ~1e-04,beta is 0.08 and mode age is 83.498"
## [1] "In 1989 , The estimated alpha is ~1e-04,beta is 0.08 and mode age is 83.826"
## [1] "In 1990 , The estimated alpha is ~9e-05,beta is 0.08 and mode age is 84.104"
## [1] "In 1991 , The estimated alpha is ~9e-05,beta is 0.081 and mode age is 84"
## [1] "In 1992 , The estimated alpha is ~9e-05,beta is 0.08 and mode age is 84.15"
## [1] "In 1993 , The estimated alpha is ~9e-05,beta is 0.08 and mode age is 84.132"
## [1] "In 1994 , The estimated alpha is ~9e-05,beta is 0.081 and mode age is 84.146"
## [1] "In 1995 , The estimated alpha is ~8e-05,beta is 0.082 and mode age is 84.33"
## [1] "In 1996 , The estimated alpha is ~8e-05,beta is 0.083 and mode age is 84.564"
## [1] "In 1997 , The estimated alpha is ~7e-05,beta is 0.084 and mode age is 84.736"
## [1] "In 1998 , The estimated alpha is ~7e-05,beta is 0.084 and mode age is 84.921"
## [1] "In 1999 , The estimated alpha is ~6e-05,beta is 0.085 and mode age is 84.879"
## [1] "In 2000 , The estimated alpha is ~6e-05,beta is 0.084 and mode age is 85.217"
## [1] "In 2001 , The estimated alpha is ~7e-05,beta is 0.084 and mode age is 85.442"
## [1] "In 2002 , The estimated alpha is ~6e-05,beta is 0.084 and mode age is 85.643"
## [1] "In 2003 , The estimated alpha is ~6e-05,beta is 0.085 and mode age is 85.665"
## [1] "In 2004 , The estimated alpha is ~6e-05,beta is 0.084 and mode age is 86.154"
## [1] "In 2005 , The estimated alpha is ~6e-05,beta is 0.085 and mode age is 86.044"
## [1] "In 2006 , The estimated alpha is ~6e-05,beta is 0.085 and mode age is 86.472"
## [1] "In 2007 , The estimated alpha is ~6e-05,beta is 0.085 and mode age is 86.49"
## [1] "In 2008 , The estimated alpha is ~5e-05,beta is 0.086 and mode age is 86.745"
## [1] "In 2009 , The estimated alpha is ~5e-05,beta is 0.085 and mode age is 87.032"
## [1] "In 2010 , The estimated alpha is ~5e-05,beta is 0.086 and mode age is 87.178"
## [1] "In 2011 , The estimated alpha is ~5e-05,beta is 0.085 and mode age is 87.61"
```

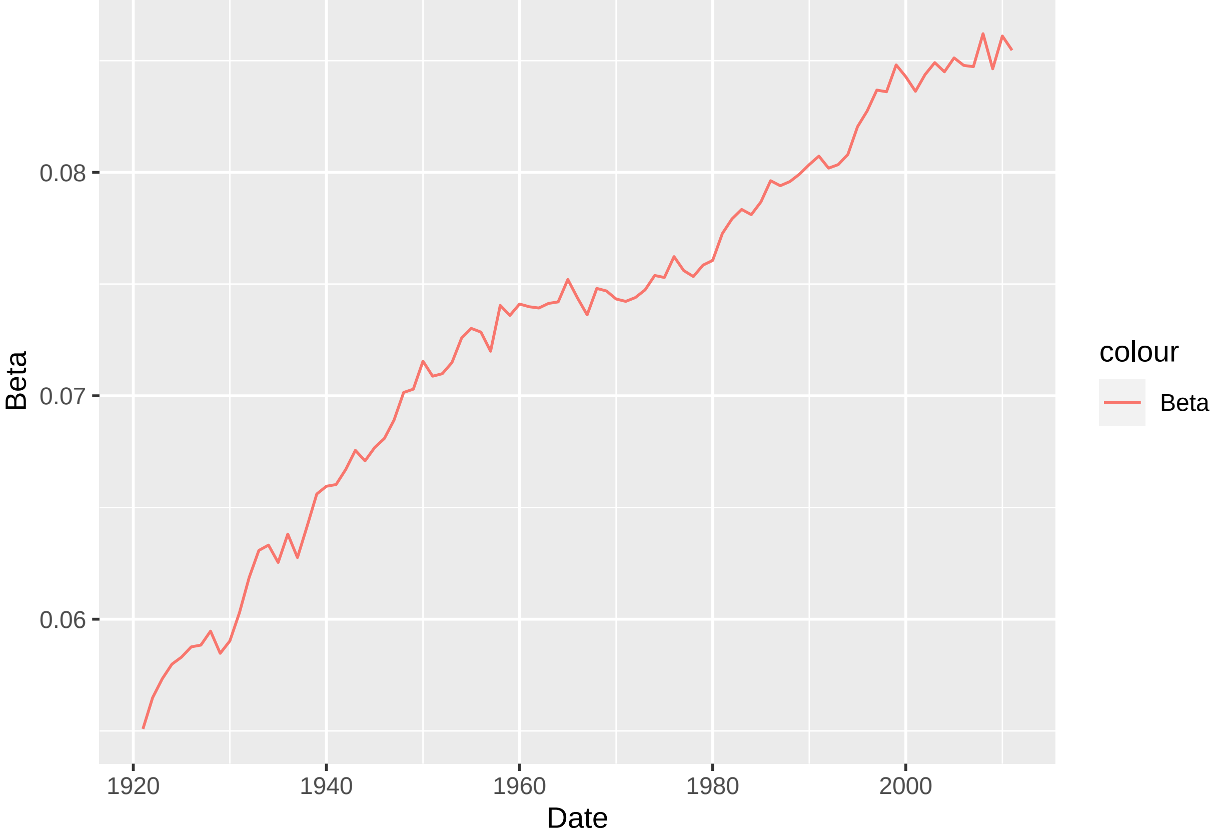
c) Create plots of α over time, β over time and the mode age at death over time. Write a few sentences interpreting these results in terms of how mortality has changed over time.

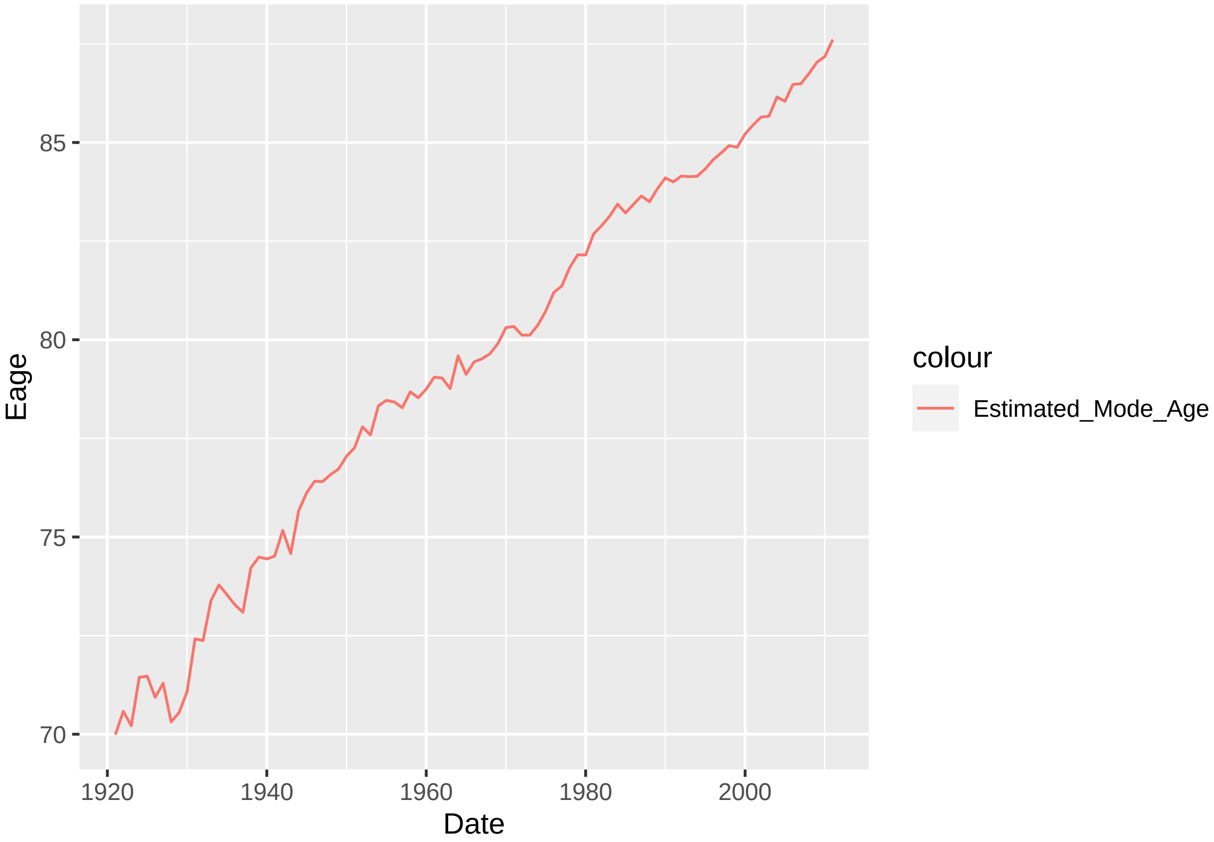
Answer :

From the plots we could see Alpha decrease exponentially as the time increase. For Beta and the Estimated Mode Age increase linearly as the time increase.

We could say the mortality keeps decrease as the time passes. The reason is the mode age of death keeps increase.


0.0012
0.0009
0.0006
0.0003
0.0000
Alpha
1920
1940
1960
1980
2000
Date
colour
Alpha


0.08
0.07
0.06
Beta
1920
1940
1960
1980
2000
Date
colour
Beta


Eage
85
80
75
70
1920
1940
1960
1980
2000
Date
colour
Estimated_Mode_Age

4 Infant mortality

In this part we will be looking at the infant mortality data set. This is in the `data` folder called `infant.RDS`.This dataset contains individual-level data (i.e., every row is a death) on deaths in the first year of life for the US 2012 birth cohort. A second dataset you will be using for this question is `births.RDS`, which tabulates the total number of live births for the US 2012 birth cohort by race and prematurity. Descriptions of each variable can be found in the `infant_mortality_codebook.txt` file.

The goal is to investigate differences in ages at death by race of mother and prematurity (from extremely preterm to full-term).

a) The infant mortality rate (IMR) is defined as the number of deaths in the first year divided by the number of live births. Calculate the IMR for the non-Hispanic black (NHB) and non-Hispanic white (NHW) populations. What is the ratio of black-to-white mortality?

Answer:

The NHB IMR is about 0.11 and the NHW IMR is about 0.005, the ratio of black to white mortality will be about 2.21.

```
## [1] 0.0109975
```

```
## [1] 0.004978799
```

```
## [1] 2.208866
```

b) Calculate the Kaplan-Meier estimate of the survival function for each race and prematurity category (i.e. you should end up with 8 sets of survival functions). Also calculate the standard error of the estimates of the survival function. Note that to calculate the survival function you will need to incorporate information from the births file, not just the deaths (otherwise it will look like everyone died).

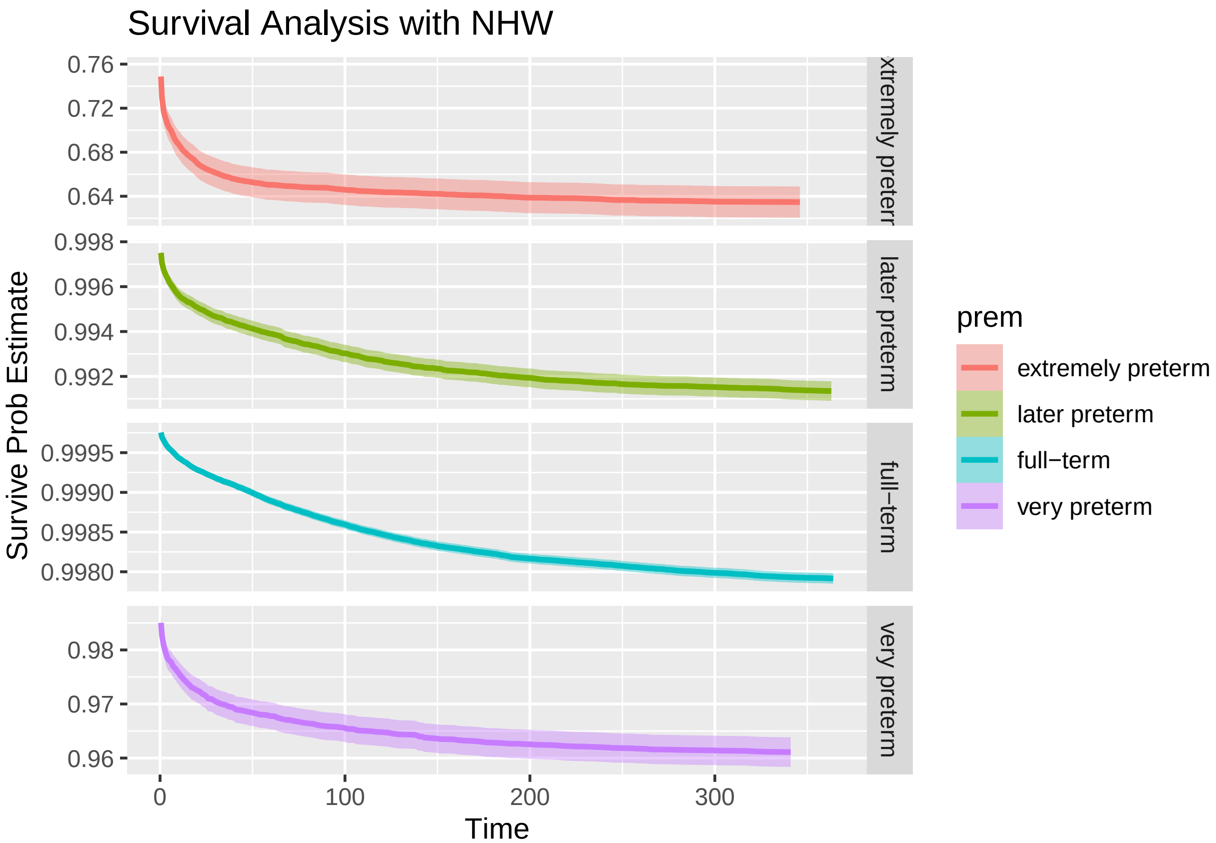
Answer;

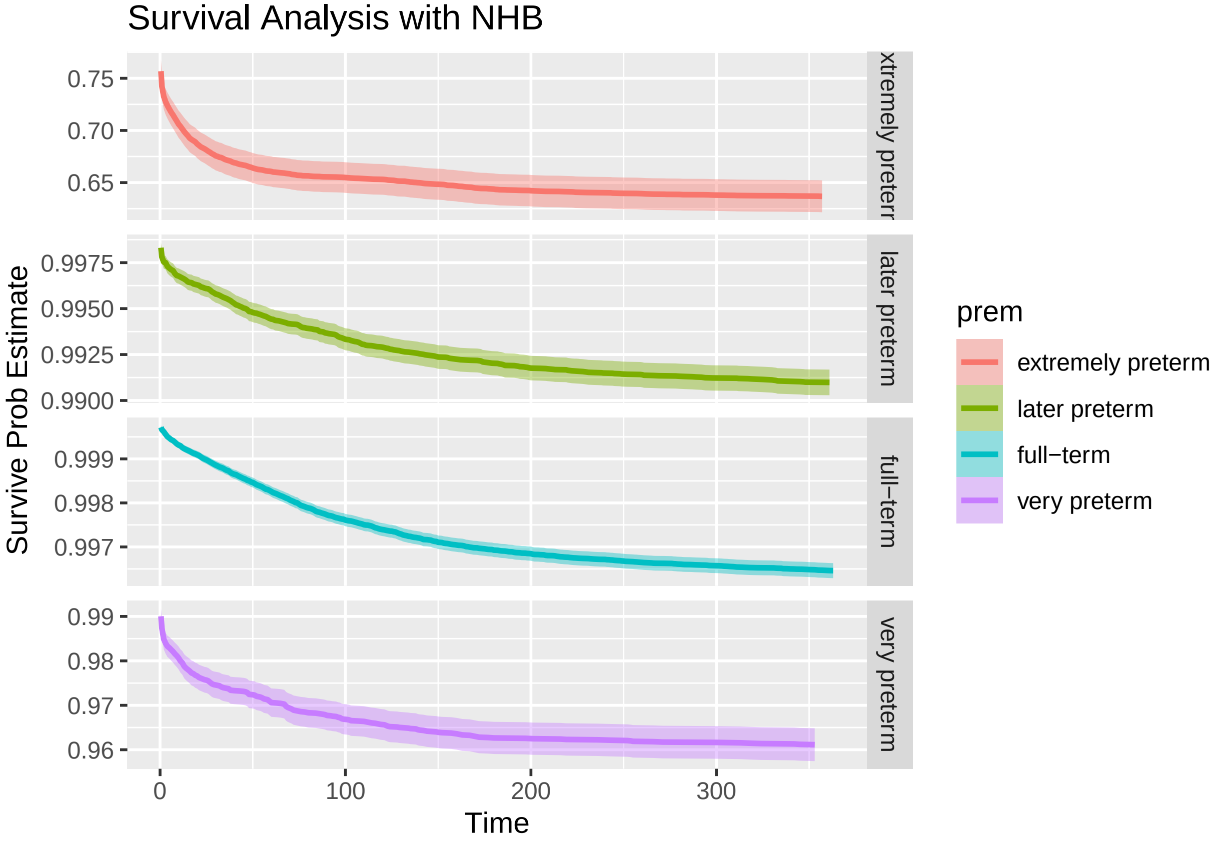
Since the table is too large for display. I have hide the code and the estimations. More details are in Rmd file and the visualization will be in part (c)

c) Plot your results from b), showing the estimate and +/- 2 standard errors. What the plot should look like: NHB and NHW survival curves on the one plot; one separate facet per prematurity category. Note that the survival curves are very different by prematurity category, so it might help to make the y axes different scales for each category (e.g. `facet_grid(prematurity~., scales = "free_y")`).

Answer:

The visulizations could be viewed below:





d) On first glance, your plots in c) might contradict what you expected based on a). Why is the

IMR so much higher for the NHB population, even though for (most) prematurity groups, the survival curves are reasonably similar to the NHW population?

Answer:

The IMR ratio are measured as a total. However, the plots are measured in time. So across different time slot there may be a little difference. But stack such effects may cause a significant different in the final IMR. By comparing the two visualizations we could conclude NHW has slightly higher survival rate than NHB across all four categories.

e) Now consider fitting a piece-wise constant hazards model to the survival time data with cut-points at 1, 7, 14, 28, 60, 90 and 120 days. Consider a model that has race and prematurity as covariates. You *could* fit this model just using the deaths data, but the direction of the sign of the coefficient on race would be misleading. Why is that?

Answer:

We could fit the model using only death data. Since we already know NHW has lower IMR ratio so the sign here is not reasonable for sure. I guess the reason might be exclusion of the birth data. By doing that we exclude the effect of total population. There might be more infants in NHW than NHB so the death speed might be higher. By summary the data we could confirm the result. NHW has 10617 and NHB only have 6407 people.

```
##
## Call:
## glm(formula = status ~ offset(log(interval_length)) - 1 + interval +
##     race + prematurity, family = "poisson", data = PW_split)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8914  -0.7379  -0.5074   0.6299   2.8400
##
## Coefficients:
##                          Estimate Std. Error  z value Pr(>|z|)
## interval0                -0.12703    0.01574   -8.072 6.90e-16 ***
## interval1                -2.79317    0.02799  -99.790  < 2e-16 ***
## interval7                -3.24692    0.03467  -93.651  < 2e-16 ***
## interval14               -3.56662    0.03260 -109.409  < 2e-16 ***
## interval28               -3.74592    0.02881 -130.001  < 2e-16 ***
## interval60               -3.78709    0.03431 -110.387  < 2e-16 ***
## interval90               -3.78134    0.03835  -98.594  < 2e-16 ***
## interval120              -3.57049    0.02549 -140.086  < 2e-16 ***
## raceNHW                   0.11662    0.01608    7.251 4.14e-13 ***
## prematurityvery preterm  -0.63380    0.03020  -20.984  < 2e-16 ***
## prematuritylater preterm -0.93639    0.02448  -38.244  < 2e-16 ***
## prematurityfull-term     -1.23917    0.01880  -65.907  < 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: 1424154  on 56513  degrees of freedom
## Residual deviance:   51073  on 56501  degrees of freedom
## AIC: 85145
##
## Number of Fisher Scoring iterations: 6
```

f) Fit a piece-wise constant hazards model with cut-points as specified in e). Note given the large numbers of births/deaths, it will be much easier to run the model based on the tabulated deaths/exposures by age at death, rather than individual-level data. Include as covariates race and prematurity, and allow the hazard ratios of each to vary by interval. Note that you may want to investigate interaction terms. Calculate the hazard of dying in the first interval (0-1 day) of extremely preterm babies born to NHB mothers. In addition, give the hazard ratios of dying for:

   1) extremeley preterm babies to NHW mothers compared to extremeley preterm babies to NHB mothers in the first interval (0-1 days).
   2) full-term babies to NHB mothers compared to extremeley preterm babies to NHB mothers in the first interval (0-1 days).
   3) full-term babies to NHB mothers compared to extremeley preterm babies to NHB mothers in the last interval (120-365 days).
   4) full-term babies to NHW mothers compared to full-term babies to NHB mothers in the last interval (120-365 days).

Answer: To better answer this question, we need to calculate the harzard of the following conditions, the:

- (0-1 day) of extremely preterm babies born to NHW mothers : 1.125
- (0-1 day) of extremely preterm babies born to NHB mothers : 1.067
- (0-1 day) of full preterm babies born to NHB mothers : 0.0998
- (120-365 day) of full preterm babies born to NHB mothers : 0.0123
- (120-365 day) of extremely preterm babies born to NHB mothers : 0.0139
- (120-365 day) of full preterm babies born to NHW mothers : 0.0113

Then we could answer the main question: The hazard of dying in the first interval (0-1 day) of extremely preterm babies born to NHB mothers is ~1.067. Then for the following question:

1. 1.125/1.067 = 1.054
2. 0.0998/1.067 = 0.0935
3. 0.0123/0.0139 = 0.884
4. 0.0113/0.0123 = 0.918

```
##
## Call:
## glm(formula = status ~ offset(log(interval_length)) - 1 + interval,
##     family = "poisson", data = PW_split)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2273  -0.5554  -0.5030  -0.4127   2.7466
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)    
## interval0   -1.86941    0.04156  -44.98   <2e-16 ***
## interval1   -3.97246    0.05234  -75.89   <2e-16 ***
## interval7   -4.40912    0.06337  -69.58   <2e-16 ***
## interval14  -4.70673    0.05488  -85.76   <2e-16 ***
## interval28  -4.76341    0.04096 -116.29   <2e-16 ***
## interval60  -4.71903    0.04762  -99.10   <2e-16 ***
## interval90  -4.63886    0.05263  -88.14   <2e-16 ***
## interval120 -4.47718    0.03061 -146.25   <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: 636457  on 20073  degrees of freedom
## Residual deviance:  15944  on 20065  degrees of freedom
## AIC: 23940
## 
## Number of Fisher Scoring iterations: 6
```

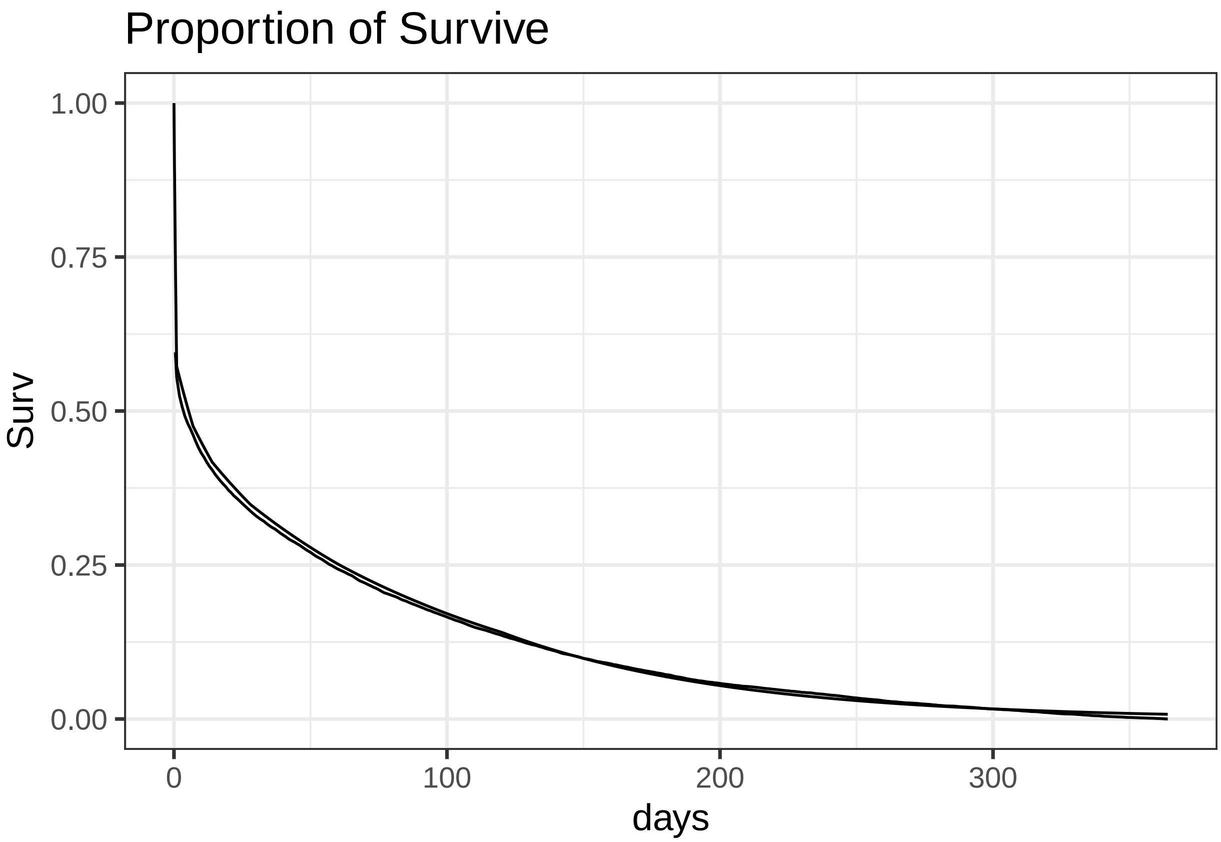
```
##   interval0   interval1   interval7  interval14  interval28  interval60 
## 0.154214942 0.018827049 0.012165926 0.009034259 0.008536481 0.008923874 
##  interval90 interval120 
## 0.009668693 0.011365452 
```

```
## [1] extremely preterm later preterm     full-term         very preterm     
## Levels: extremely preterm very preterm later preterm full-term
```

g) Fit a piecewise hazards model to the whole population (i.e. just have `interval` as a covariate) and calculate the survival curve. Compare to the KM estimate from b) by plotting the two curves on the one graph. The fit should be fairly reasonable, so if it's not there could be an issue in your part f) model.

Answer:

Since in this question we mainly use the deaths data. So for the KM curve we will keep use deaths data. We could see the curve fit almost the same.


Proportion of Survive
Surv
1.00
0.75
0.50
0.25
0.00
0
100
200
300
days