STATISTICAL RISK MODELLING

SEMESTER-3 PROJECT REPORT

- PALAK NAGDEV (419)

1)

After setting the working directory, we read the data using read.table (using "," and keeping header as "TRUE")

```
data=read.table('Graduation(1).csv' , header=TRUE , sep=',')
data$CRUDE<-data$DEATHS/data$ETR
```

Checking head and tail of the data:

```
head(data)
tail(data)
```

```
> head(data)
  AGE   ETR DEATHS        CRUDE GRADUATED EXPECTED    ZX
1  25 78500     24 0.0003057325  0.000238    18.68  1.23
2  26 80425     24 0.0002984147  0.000265    21.31  0.58
3  27 81975     24 0.0002927722  0.000294    24.10 -0.02
4  28 83725     24 0.0002866527  0.000327    27.38 -0.65
5  29 84875     72 0.0008483063  0.000364    30.89  7.40
6  30 85075     48 0.0005642081  0.000405    34.46  2.31
> tail(data)
   AGE    ETR DEATHS      CRUDE GRADUATED EXPECTED    ZX
46  70 129225   4428 0.03426582  0.028438  3674.90 12.42
47  71 128875   3915 0.03037827  0.031627  4075.93 -2.52
48  72 130075   5103 0.03923121  0.035174  4575.26  7.80
49  73 130475   5454 0.04180111  0.039119  5104.05  4.90
50  74 129550   6453 0.04981088  0.043507  5636.33 10.88
51  75 129400   6453 0.04986862  0.048386  6261.15  2.42
> 
```

2)

Filling the graduated column using gompertz law, finding parameters B and C by using the functions "coef" and "as.numeric" and Using round function to rounding off the graduated column to 5 decimal palces.

B = 1.668727e-05

C = 1.112153e+00

```
gompertz1<-lm(log(data$CRUDE)~data$AGE)
gompertz1
coef(1)
B=exp(as.numeric(coef(gompertz1)))[1]
C=exp(as.numeric(coef(gompertz1)))[2]
c(B,C)
data$GRADUATED<-round(B*C^data$AGE,5)
```

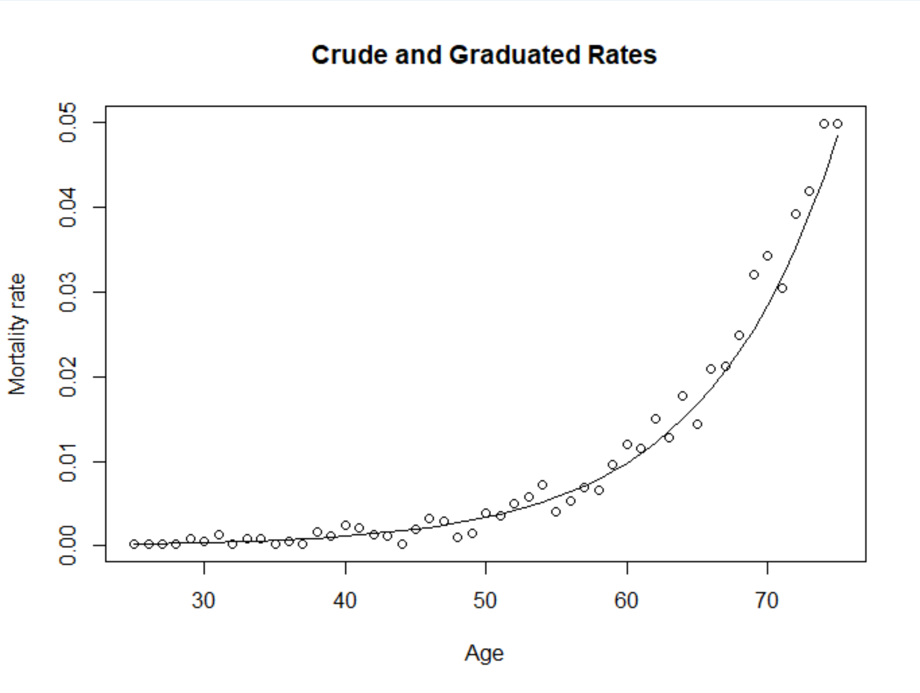
Results:

```
Call:
lm(formula = log(data$CRUDE) ~ data$AGE)

Coefficients:
(Intercept)     data$AGE
   -11.0009       0.1063
```

Visualizing:

```
plot(data$AGE,data$CRUDE, xlab="Age",ylab="Mortality rate",main="Crude and Graduated Rates")
lines(data$AGE,data$GRADUATED)
```



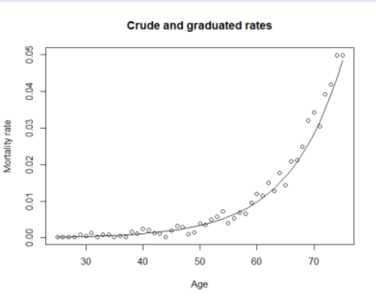
3) Checking for smoothness by applying 3rd differences to graduated and crude rates using the diff function.

3rd differences should be small in magnitude and progress regularly to check smoothness. In our data frame the 3rd differences of crude rates are much larger in magnitude and progresses erratically.However the magnitude of 3rd differences of the graduated rates are very small and progress regularly.

Therefore, the desired graduated rates computed by fitting Gompertz Law are smooth.

```
diff_new<-function(v)v[-1]-v[-length(v)]
diff_crude=round(diff_new(diff_new(diff_new(data$CRUDE)))*10^6,0)
diff_grad=round(diff_new(diff_new(diff_new(data$GRADUATED)))*10^6,0)
plot(data$AGE,data$CRUDE, xlab="Age",ylab="Mortality rate",main="Crude and graduated rates")
lines(data$AGE,data$GRADUATED)
cbind(data$AGE[data$AGE<=72],diff_crude,diff_grad)
```

Visualizing :




```
> cbind(data$AGE[data$AGE<=72],diff_crude,diff_grad)
        diff_crude diff_grad
 [1,] 25         -2         0
 [2,] 26        568       -20
 [3,] 27      -1414        20
 [4,] 28       1973         0
 [5,] 29      -3099       -10
 [6,] 30       3648        10
 [7,] 31      -2233       -10
 [8,] 32         17        10
 [9,] 33       1342         0
[10,] 34      -1322       -10
[11,] 35       2241        20
[12,] 36      -3676       -20
[13,] 37       3676        20
[14,] 38      -3183       -10
[15,] 39        947        10
[16,] 40       1004       -10
[17,] 41      -1142        10
[18,] 42       3333         0
[19,] 43      -3124        10
[20,] 44      -1295       -10
[21,] 45        398         0
[22,] 46       3528        20
[23,] 47       -274       -10
[24,] 48      -4533        10
[25,] 49       4447        10
[26,] 50      -2588       -10
[27,] 51       1433        10
[28,] 52      -5286        10
[29,] 53       9026        10
[30,] 54      -4183         0
[31,] 55      -1951        10
[32,] 56       4992        10
[33,] 57      -3863        10
[34,] 58      -2369        10
[35,] 59       6962        20
[36,] 60      -9542         0
[37,] 61      12421        30
[38,] 62     -14898         0
[39,] 63      17650        40
[40,] 64     -15583         0
[41,] 65       9139        40
[42,] 66        267        20
[43,] 67      -8292        30
[44,] 68      -1303        30
[45,] 69      18882        30
[46,] 70     -19024        60
[47,] 71      11723        30
[48,] 72     -13392        50
```

4)

Calculating expected rates and Zx and rounding both the numbers to 3 decimal places. Conducting a chi-square test using chisq.test function.

Interpretation :

➔ Graduation is not Okay

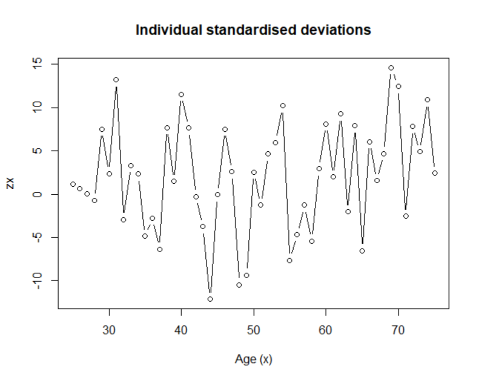
H0: Graduation is Okay

H1: Graduation is not Okay

Since the p-value is greater than 0.05, H0 is rejected.

```
data$EXPECTED<-round(data$GRADUATED*data$ETR,3)
data$ZX<-round((data$DEATHS-data$EXPECTED)/sqrt(data$EXPECTED),3)
plot(data$AGE,data$ZX,type="b",xlab="Age (x)", ylab="zx",main="Individual standardised deviations")
data$prob = data$EXPECTED/sum(data$EXPECTED)
c=chisq.test(data$DEATHS,data$prob,rescale.p = TRUE)
c$parameter
c$p.value
c$statistic
sum(data$ZX^2)
```

Visualizing:



5)

a)

STANDARDISED DEVIATIONS TEST

Creating intervals of 1 from -3 to +3 under binning_var 1. Creating a testing table and a probability table using prop.table function.

Finally, creating a new dataframe with observed and expected probability in the columns.

Conducting a χ^2test using chisq.test function.

OVERALL SHAPE - The data shows that it is negatively skewed and has more values on the tail.

ABSOLUTE DEVIATIONS – The data is over graduated and shows existence of duplicates as the absolute deviations are too big (shown by proved by the fact that there are less than 50% values lying in the (-2/3 to 2/3) bin)

OUTLIERS – Too many outliers are present as close to 20% of the data lies in the (-infinity to -3) bin and 40% of the data lies in the (3 to Infinity) bin. Showing that the data has approximately 60% outliers.

SYMMETRY – The number of positive and the number of negative deviations are not close to 50% rather there are only 20 negative and 31 positive deviations. This shows that observed mortality rates do not conform to the model with the rates assumed in the graduation. The data shows nature of discrepancy.

```
binning_var1 = c(-Inf, -3, -2, -1, 0, 1, 2, 3, Inf)
testing_table1 = data.frame(data$ZX, bin=cut(data$ZX, binning_var1, include.lowest=TRUE))
prop.table(table(testing_table1$bin))
standdev_test = data.frame(Bins = levels(unique(testing_table1$bin)), Expected = c(0, 0.02, 0.14, 0.34, 0.34, 0.14, 0.02, 0), Observed = round(as.r
standdev_test = standdev_test %>% mutate(Expected = Expected * 51, Observed = Observed * 51)
chisq.test(standdev_test$Expected, standdev_test$Observerd, correct = T)
```

```
> prop.table(table(testing_table1$bin))

 [-Inf,-3]    (-3,-2]    (-2,-1]     (-1,0]      (0,1]      (1,2] 
0.19607843 0.07843137 0.03921569 0.05882353 0.03921569 0.07843137 
     (2,3]   (3, Inf] 
0.11764706 0.39215686 
```

```
	Chi-squared test for given probabilities

data:  standdev_test$Expected
X-squared = 59.65, df = 7, p-value = 1.773e-10
```

H0: Graduation is Okay

H1: Graduation is not Okay

χ^2 test gives a p.value of less than 0.05 and as mentioned above as the p.value was very low we have insufficient evidence to reject H0 and therefore the Graduation is OK.

```
data$signs = ifelse(data$GRADUATED>data$CRUDE,1,0)
head(data)
tail(data)
sum(data$signs)
p=2*pbinom(20,51,0.5)
p
```

Signs test checks for defect (b) of χ^2 test.

H0 : Data is biased

H1 : Data is not biased

As the p.value = 0.16 which is greater than 0.05 we have insufficient evidence to reject H0 . Hence data is not biased.

CUMULATIVE DEVIATIONS TEST

```
(sum(data$DEATHS)-sum(data$EXPECTED))/(sqrt(sum(data$EXPECTED)))


#5)d

z1=data$ZX[1:length(data$ZX)-1]
z2=data$ZX[2:length(data$ZX)]
a=cor(z1,z2)
a
a*sqrt(51)
```

It checks cumulations of positive and negative groups.

Test statistic = 18.831

H0: Graduation rates are OK

H1: Graduation rates are too low

As test statistic is greater than 1.96 we have sufficient evidence to reject H0 and conclude that Graduation rates are too low.

"z1=data$ZX[1:length(data$ZX)-1]

z2=data$ZX[2:length(data$ZX)]

a=cor(z1,z2)

a

a*sqrt(51)"

SERIAL CORRELATIONS TEST

It detects clumping of signs of deviations

As a= r_j =0.1477 we can conclude that Z_x has similar values.

H0 : No grouping of signs

H1 : Grouping of signs

As test statistic is 1.05 which is less than 1.649 we have insufficient evidence to reject H0. Therefore there is no evidence of grouping of signs .