SRM Project

Name: Ishika Shah Roll No.: 422

Q1. Read in the data file as a data table & fill in the entries for the CRUDE column in your table.

```
data = read.csv(file.choose())
data
str(data)
dim(data)
data$CRUDE = data$DEATHS/data$ETR
data
```

Q2) Use Gompertz law to fill entries in the GRADUATED column in the data.

```
gl = lm(log(CRUDE)~AGE, data)
summary(gl)

B = exp(coef(gl))[1]
C = exp(coef(gl))[2]

data$GRADUATED = round(B*C^data$AGE, 6)
data
```

```
> gl = lm(log(CRUDE)~AGE, data)
> summary(gl)

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.07946 -0.09831  0.08351  0.21582  1.13949

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.000864   0.256884  -42.82   <2e-16 ***
AGE           0.106298   0.004929   21.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5181 on 49 degrees of freedom
Multiple R-squared:  0.9047,	Adjusted R-squared:  0.9028
F-statistic: 465.2 on 1 and 49 DF,  p-value: < 2.2e-16
```

Q3) Check for smoothness by applying the third differences to the crude and graduated rates and comment on your results.

```
diff1 = function(x)x[-1]-x[-length(x)]

crude_diff = round(diff1(diff1(diff1(data$CRUDE)))*10^6,0)
crude_diff
crude_data = round(diff1(diff1(diff1(data$GRADUATED)))*10^6,0)
crude_data

third_diff = cbind(data$AGE[data$AGE<=72], crude_diff, crude_data)
third_diff
```

```
> diff1 = function(x)x[-1]-x[-length(x)]
> crude_diff = round(diff1(diff1(diff1(data$CRUDE)))*10^6,0)
> crude_diff
 [1]     -2    568  -1414   1973  -3099   3648  -2233     17   1342  -1322   2241  -3676   3676  -3183    947   1004  -1142   3333  -3124  -1295    398
[22]   3528   -274  -4533   4447  -2588   1433  -5286   9026  -4183  -1951   4992  -3863  -2369   6962  -9542  12421 -14898  17650 -15583   9139    267
[43]  -8292  -1303  18882 -19024  11723 -13392
> crude_data = round(diff1(diff1(diff1(data$GRADUATED)))*10^6,0)
> crude_data
 [1]  2  0  0  0  2 -1  1  2 -1  2  1  0  2  2  0  2  3  1  3  2  2  5  2  5  3  6  4  7  7  6  9  9  9 13 11 15 14 19 17 23 23 25 31 31 37 40 45 48
> third_diff = cbind(data$AGE[data$AGE<=72], crude_diff, crude_data)
> third_diff
         crude_diff crude_data
 [1,] 25         -2          2
 [2,] 26        568          0
 [3,] 27      -1414          0
 [4,] 28       1973          0
 [5,] 29      -3099          2
 [6,] 30       3648         -1
 [7,] 31      -2233          1
 [8,] 32         17          2
 [9,] 33       1342         -1
```

The third differences of the crude rates are much greater in magnitude than the graduated rates.

We can see that from the table that the third differences of the crude rates are much larger in magnitude (max = 19024) and progress is irrational.

This is to be expected, since they are calculated directly from the deaths, which include significant random elements.

The third differences of the graduated rates are all very small and progress regularly.

This is to be expected, since they have been smoothened using a simple parametric formula with just two parameters.

Q4) Calculate the values in EXPECTED and ZX values in the table. Hence, perform a chi-squared test to check goodness of fit between DEATHS and EXPECTED. You should specify the degrees of freedom used.

```
data$EXPECTED = round(data$ETR*data$GRADUATED,2)
data$ZX = round((data$DEATHS-data$EXPECTED)/sqrt(data$EXPECTED),3)
data

chisq = data.frame(data$DEATHS, data$EXPECTED)
chisq
chisq.test(chisq)
```

```
> data
```

	AGE	ETR	DEATHS	CRUDE	GRADUATED	EXPECTED	ZX
1	25	78500	24	0.0003057325	0.000238	18.68	1.231
2	26	80425	24	0.0002984147	0.000265	21.31	0.583
3	27	81975	24	0.0002927722	0.000294	24.10	-0.020
4	28	83725	24	0.0002866527	0.000327	27.38	-0.646
5	29	84875	72	0.0008483063	0.000364	30.89	7.397
6	30	85075	48	0.0005642081	0.000405	34.46	2.307
7	31	85275	120	0.0014072120	0.000450	38.37	13.178
8	32	86250	24	0.0002782609	0.000501	43.21	-2.922
9	33	87250	72	0.0008252149	0.000557	48.60	3.357
10	34	88300	72	0.0008154020	0.000619	54.66	2.345
11	35	90200	24	0.0002660754	0.000689	62.15	-4.839
12	36	92500	48	0.0005189189	0.000766	70.86	-2.716
13	37	95425	24	0.0002515064	0.000852	81.30	-6.355
14	38	98550	168	0.0017047184	0.000948	93.43	7.715
15	39	99775	120	0.0012027061	0.001054	105.16	1.447
16	40	99125	240	0.0024211854	0.001172	116.17	11.489
17	41	99200	216	0.0021774194	0.001304	129.36	7.618
18	42	101525	144	0.0014183699	0.001450	147.21	-0.265
19	43	104525	120	0.0011480507	0.001612	168.49	-3.736
20	44	107075	24	0.0002241420	0.001793	191.99	-12.124
21	45	109125	216	0.0019793814	0.001994	217.60	-0.108
22	46	109425	360	0.0032899246	0.002218	242.70	7.529
23	47	109075	312	0.0028604171	0.002467	269.09	2.616
24	48	110175	120	0.0010891763	0.002743	302.21	-10.481
25	49	111675	168	0.0015043653	0.003051	340.72	-9.357
26	50	112725	432	0.0038323353	0.003393	382.48	2.532
27	51	115250	408	0.0035401302	0.003774	434.95	-1.292
28	52	118225	600	0.0050750687	0.004197	496.19	4.660
29	53	120025	702	0.0058487815	0.004668	560.28	5.987
30	54	122150	891	0.0072943103	0.005191	634.08	10.203
31	55	124350	513	0.0041254524	0.005773	717.87	-7.646
32	56	125750	675	0.0053677932	0.006421	807.44	-4.661
33	57	126350	864	0.0068381480	0.007141	902.27	-1.274
34	58	127100	837	0.0065853659	0.007942	1009.43	-5.427
35	59	129350	1242	0.0096018554	0.008833	1142.55	2.942
36	60	132475	1593	0.0120249104	0.009823	1301.30	8.086
37	61	134000	1539	0.0114850746	0.010925	1463.95	1.961
38	62	133700	1998	0.0149439043	0.012150	1624.45	9.268
39	63	134375	1728	0.0128595349	0.013513	1815.81	-2.061
40	64	136125	2403	0.0176528926	0.015028	2045.69	7.900
41	65	136625	1971	0.0144263495	0.016714	2283.55	-6.541
42	66	136100	2835	0.0208302719	0.018588	2529.83	6.067
43	67	135750	2889	0.0212817680	0.020673	2806.36	1.560
44	68	134350	3348	0.0249199851	0.022992	3088.98	4.660
45	69	131575	4212	0.0320121604	0.025570	3364.37	14.614
46	70	129225	4428	0.0342658154	0.028438	3674.90	12.423
47	71	128875	3915	0.0303782735	0.031627	4075.93	-2.521
48	72	130075	5103	0.0392312128	0.035174	4575.26	7.802
49	73	130475	5454	0.0418011113	0.039119	5104.05	4.898
50	74	129550	6453	0.0498108838	0.043507	5636.33	10.878
51	75	129400	6453	0.0498686244	0.048386	6261.15	2.425

```
        Pearson's Chi-squared test

data:  chisq
X-squared = 905.22, df = 50, p-value < 2.2e-16
```

Null hypothesis: The graduated rates are the true underlying mortality rates

Alternate hypothesis: Mortality Rate is not consistent with the set of Graduated Mortality Rates

DOF = 50

Since p value < 0.05, we reject the null hypothesis at 5% level of significance, and conclude that the graduated rates are not in line with the actual mortality rates.

Q5)

a. Perform the standardised deviations test on the individual deviations, and comment on the-

i. Overall shape
ii. Absolute deviations
iii. Outliers
iv. Symmetry
v. Final conclusion about Null hypothesis

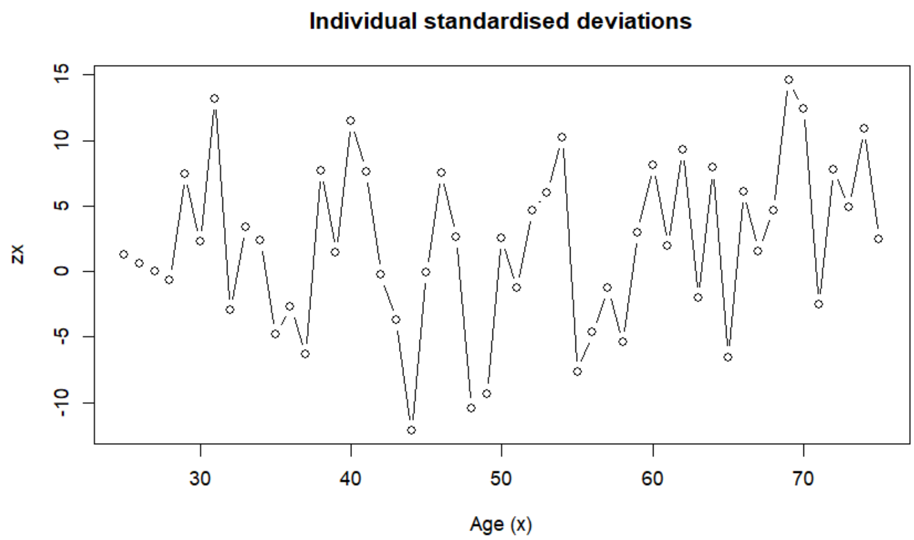
b. Perform the Signs test and give your conclusion
(Hint : Since m is large, you may use the normal approximation along with continuity correction for this test. You may look up the function *qnorm* and *pnorm* in R for this)
c. Perform the Cumulative deviations test for the entire age range and give your conclusion.
d. Perform the Serial correlations test and give your conclusion

```
library(plyr)
install.packages("EnvStats")
library(EnvStats)
plot(data$AGE,data$ZX,type="b",xlab="Age (x)", ylab="zx",main="Individual standardised deviations")
table(cut(data$ZX, breaks = seq.int(from = -20,to = 20, by= 4)))
min(data$ZX)
max(data$ZX)
iqr(data$ZX)
boxplot(data$ZX)$out
skewness(data$ZX)
```

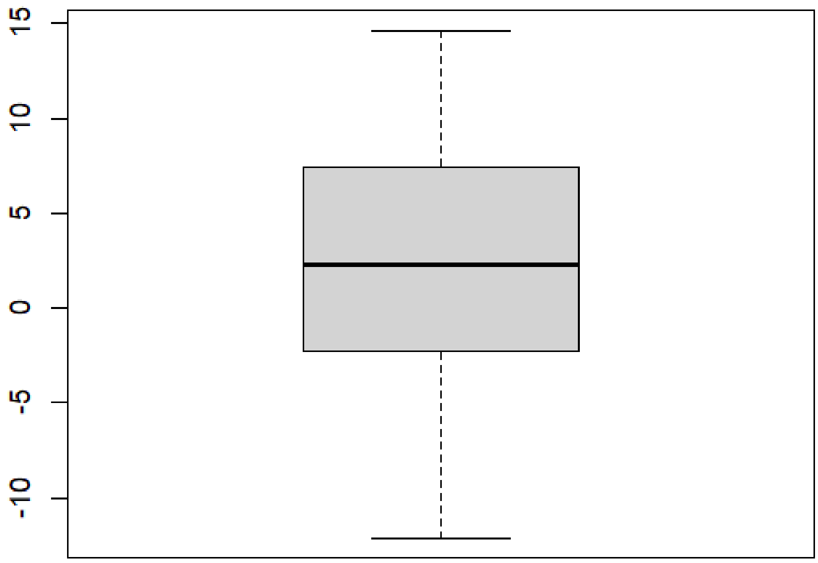
a) Individualised standard deviations test

```
> table(cut(data$ZX, breaks = seq.int(from = -20,to = 20, by= 4)))

(-20,-16] (-16,-12]  (-12,-8]   (-8,-4]    (-4,0]     (0,4]     (4,8]    (8,12]   (12,16]   (16,20]
        0         1         2         6        11        12        11         5         3         0
```



```
> min(data$ZX)
[1] -12.124
> max(data$ZX)
[1] 14.614
> iqr(data$ZX)
[1] 9.754
> boxplot(data$ZX)$out
numeric(0)
> skewness(data$ZX)
[1] -0.1311715
```



(i) The graph is much wider a standard normal graph, with several values higher than 10.
(ii) The values of the absolute deviations are much higher relative to the expected value.
(iii) The lower bound is -12.124 and the upper bound is 14.614. IQR is 9.754. There are no outliers.
(iv) The Graph is negatively skewed.
(v) The graduated rates do not represent the underlying mortality rates with accuracy.

b) Signs test

```
> signs_test = sign(data$ZX)
> table(signs_test)
signs_test
-1  1
20 31
> pbinom(31,51,0.5)
[1] 0.9540427
```

H0: There is no bias in the data.

H1: There is bias in the data

We fail to reject the null hypothesis that the data is a true representation of the underlying mortality rates.

c) Cumulative deviations test

```
> ## Cumulative deviations
> observed = sum(data$DEATHS)
> expected = sum(data$EXPECTED)
> z=(observed - expected)/sqrt(expected)
> z
[1] 18.83091
```

H0: Graduated rates are not biased.
H1: Graduated rates are biased.

We reject the null hypothesis and conclude that the graduated rates are biased.

d) Serial Correlations Test

```
> ## Serial correlation
> z1 = data$ZX[1:length(data$ZX)-1]
> z2 = data$ZX[2:length(data$ZX)]
> sctest = cor(z1, z2)
> sctest
[1] 0.1476081
> teststat = sqrt(51)*sctest
> teststat
[1] 1.054133
```

H0: Grouping of signs is absent.

H1: Grouping of signs is present

At 5% significance level test statistic is 1.6449

Therefore, we fail to reject the null hypothesis.