ACTL3141 – Mortality Analysis of Chilean Annuitants

Dhwanish Kshatriya

(z5421168)

11 April 2024

Contents

Section 1 – Descriptive Analysis

The following is a descriptive analysis of the mortality data used in the construction of the current life tables used in Chile. The dataset includes information on 1,292,017 annuitants aged 60 or more who were a part of the pension system between 1st January 2014 and 31st December 2018. In this descriptive analysis, we aim to briefly explore the annuitant mortality data and gain insights on the description and profile of annuitants. Certain anomalies were identified and thus data cleaning and preparation was performed. This is explained in Explanation 1.

Representation of Age & Covariates

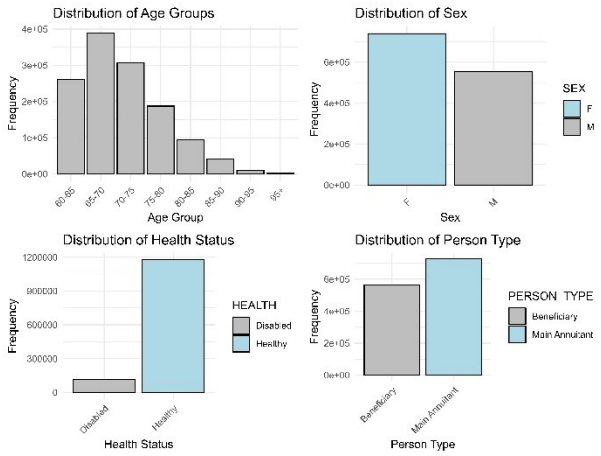

Figure 1 - Histograms of Distributions of Age & Covariates

The first histogram in Figure 1 showcases the distribution of annuitant Age at the end of exposure (by age group). We can see that the distribution of Age is highly right-skewed, with majority of annuitants falling between ages 60 to 75. We can also see a significant decline in the frequency of annuitants at higher age groups beyond 85 years, indicating a relative under-representation of these ages in the dataset. Next, looking at the covariate of HEALTH, we see that the frequency of disabled annuitants is significantly higher than that of healthy annuitants, with disabled annuitants making up only 8.8% of the dataset. Due to the size of the dataset, this 8.8% represents over 100,000 observations, which is generally large enough to have a minimal impact on our statistical estimates later. Conversely, the distribution of covariates SEX and PERSON_TYPE seems to be relatively balanced. The distribution of the variables can potentially cause issues throughout the investigation. Particularly, there is a risk of producing highly variable estimates on subsets of the data that are under-represented (such as for old, disabled annuitants).

Correlation of Covariates

Looking at Table 3 - Correlation Matrix of Covariates, we can see a relatively strong, positive correlation between covariates PERSON_TYPE and SEX (with a value of 0.579). Similarly, we can also see a moderate positive correlation between PERSON_TYPE and HEALTH, with a value of -0.267. These imply a relatively weak linear relationship between PERSON_TYPE and other covariates, with the correlation being significant enough to warrant further investigation later in the study. We can also see slight negative correlation between SEX and HEALTH with a value of -0.140, suggesting a weak, negative linear relationship between the covariates. It is important to note that correlation does not imply causation, and thus while there is some correlation between variables, we cannot draw any conclusions about an interaction effect until we perform further analysis.

Death Distribution by Age & Covariate

The values discussed in this section are summarised in Table 1 and Table 2. First, we define death proportion as the number of deaths in a group divided by the total number of observations in the group. From Table 1, we can see that the death proportion is relatively low for age groups 60-65 and 65-70, at 0.048 and 0.052 respectively. From here, we observe the death proportion increase at a slightly faster rate an increase of roughly 0.02-0.03 per age group (for groups 70-75 and 75-80). Then, we see a drastic increase in the death proportion of roughly 0.07 per age group for the remaining age groups (80-85, 88-90, 90-95 and 95+). This indicates that mortality increases at an increasing rate with Age, which aligns with our current understand. Looking at the effect of each covariate in Table 2, we can see that the death proportion is higher for male annuitants (0.107) than female annuitants (0.0601). We can also observe a slightly smaller difference in death proportion based on PERSON_TYPE, with beneficiaries having a lower death proportion (0.063) as compared to main annuitants (0.094). Finally, we can observe a much more drastic difference in the death

proportion by HEALTH status. We see that disabled annuitants (0.0148) have over double the death proportion of healthy annuitants (0.0739). Thus, our descriptive analysis indicates that mortality increases with Age at an increasing rate, mortality is higher for men than women, mortality is higher main annuitants than beneficiaries and mortality is higher for disabled annuitants than healthy annuitants. However, we cannot draw any conclusions from these insights until we investigate for any confounding effects and consider exposure to risk.

Section 2 – Survival Analysis

To perform survival analysis of the Chilean annuitant mortality data, both semi-parametric and non-parametric techniques were used. A Cox regression model was fit using a backward selection technique and an AIC selection statistic. However, the model violated key assumptions and demonstrated a poor fit, namely the proportional hazards assumption and distribution of Cox-Snell residuals. This is further discussed in Explanation 2. Then, a Kaplan-Meier (KM) estimate of the survival function was found, which relies on the assumptions of non-informative censoring and independence of lives. Due to the large size of the dataset, containing over 1.2 million observations, it is much more reasonable to rely on these assumptions. Alternatively, we also had the option of also using the Nelson-Aalen (NA) estimator to estimate the survival function, however due to the size of the dataset, this was found to be essentially equivalent to the KM estimate. Therefore, for the following survival analysis, we will primarily focus on the findings of the KM estimator, only using the Cox regression model to support the investigation when required.

Key Results & Analysis of Covariates

```
Call:
coxph(formula = Surv(time = mortality$age_at_start, time2 = mortality$age_at_end,
    event = mortality$DEATH) ~ SEX + HEALTH + PERSON_TYPE, data = mortality,
    method = "breslow")

  n= 1291958, number of events= 103827

                              coef exp(coef)  se(coef)       z Pr(>|z|)
SEXM                     0.561334  1.753009  0.008673   64.72   <2e-16 ***
HEALTHHealthy           -1.091456  0.335727  0.009152 -119.26   <2e-16 ***
PERSON_TYPEMain Annuitant -0.153264  0.857904  0.009321  -16.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                         exp(coef) exp(-coef) lower .95 upper .95
SEXM                        1.7530     0.5704    1.7235    1.7831
HEALTHHealthy               0.3357     2.9786    0.3298    0.3418
```

Figure 2 - Cox Regression Model & Results (no Interaction)

```
Concordance= 0.623  (se = 0.001 )
Likelihood ratio test= 20384  on 3 df,   p=<2e-16
Wald test            = 24511  on 3 df,   p=<2e-16
Score (logrank) test = 26924  on 3 df,   p=<2e-16
```

The estimated KM survival curve for the entire mortality dataset (top-left plot in Figure 3) Showcases various general features in accordance with our understanding of human mortality. Most simply, it shows that the survival probability decreases with Age. Looking further, we can observe the magnitude of the gradient of the tangent to the curve increasing at a higher rate as Age increases. This gradient reflects the magnitude of the rate of mortality, as $\mu_x = -S'(x) \div S(x)$, which increases at an increasing rate with Age. Finally, we can also observe that the probability of survival past extremely high ages past 105 is close to 0.

Next, looking at individual covariates in Figure 3, we can see a difference between estimated survival curves based on an individual's SEX, HEALTH, and PERSON_TYPE. As it is difficult to deduce whether these differences are significant from the plots above, we can refer to the results of our Cox regression model in Figure 2 for support. Here, the p-value is less than 2×10^{-16} for each covariate SEX, HEALTH, and PERSON_TYPE. This suggests we reject the null hypothesis (that the covariate has no effect on the hazard rate) for each covariate at a 95% confidence level. Further, the log-rank test can be used from the Cox-regression model to test the null hypothesis that there is no difference between the survival times of different covariate sub-groups. One again, the p-value is less than 2×10^{-16} meaning suggesting that we reject the null hypothesis at a 95% confidence level, suggesting that the set of covariates (SEX, HEALTH, and PERSON_TYPE) have a statistically significant effect on survival. Whilst it may seem like all 3 covariates have a statistically significant effect on the survival of an individual, we cannot come to any conclusions until we test for a confounding effect.

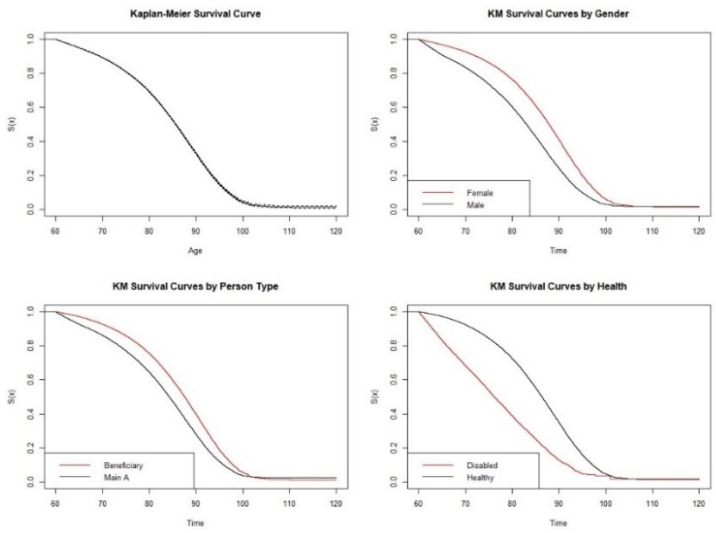

Figure 3 - KM Survival Curves by Covariate

Examining these survival curves in Figure 3, we can see various relationships between each covariate and the estimated survival function. Note that as our Cox regression model in Figure 2 violates the proportional hazard assumption, it is relatively inaccurate to draw conclusions on the exact relative hazard of annuitants from these results. Instead, we will simply look at the sign of the coefficient to see whether the covariate results in an increased or decreased hazard. Starting with HEALTH, we can see that for disabled annuitants the survival curve at an age $x > 60$ is consistently lower when compared to healthy annuitants. We can also observe that slope of the survival curve is significantly higher (in magnitude) between 60-80 years of age, suggesting that disabled annuitants have a higher rate of mortality at these ages. These insights are supported by the results of our Cox-regression model in Figure 2, which states that the coefficient for a healthy annuitant is $\beta = -1.0915 < 0$, suggesting that healthy annuitants have a lower hazard rate than a disabled annuitant ($\exp\{\beta\} < 1$ times as much). Next, looking at the estimated survival curve for SEX in Figure 3,we observe a consistently lower survival for men than women until a given age $x > 60$. We can also note that initially, the gradient of the survival curve is slightly higher (in magnitude) for men compared to women, suggesting they have a higher rate of mortality at younger ages. The insights from our Cox-regression in Figure 2 support these ideas, with the coefficient $\beta = 0.5613 > 0$, suggesting that male annuitants have a greater hazard rate than female annuitants ($\exp\{\beta\} > 1$ times as much). Finally, due to the relatively high correlation between covariates found in the descriptive analysis, we must investigate the interaction effects between PERSON_TYPE and the other covariates.

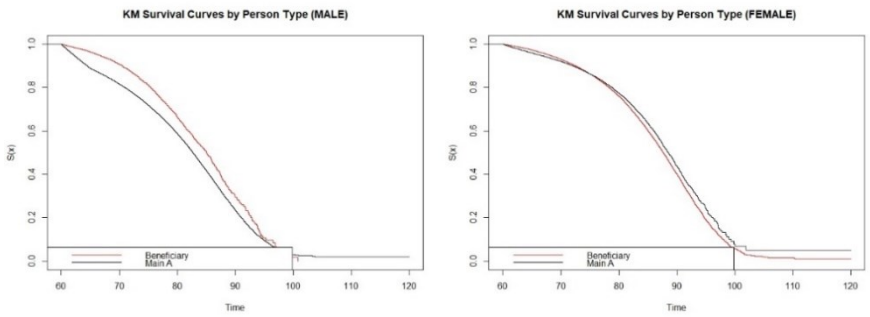

Figure 4 - KM Survival Curves by Sex & Person Type

Looking into this further, we can see in Figure 4 that the relationship between survival and PERSON_TYPE varies depending on the annuitant's SEX. The first plot shows us that for men, the survival curve is higher for beneficiaries compared to main annuitants. The second plot showcases a more complex relationship for women, where survival is initially higher for beneficiaries (up to an age of roughly 75), from where survival remains higher for main annuitants. These diverging results suggest that there is some form of interaction between PERSON_TYPE and SEX. To test this further, we constructed a Cox regression model with interaction terms between PERSON_TYPE and other covariates, the results to which are presented in Figure 7. In short, we can see that the interaction between SEX and PERSON_TYPE has a statistically significant effect on survival at a level of $\alpha = 0.05$ for females but nor for females. We can also see that the interaction between HEALTH and PERSON_TYPE is not statistically significant at this level. Looking at the coefficients in Figure 7, we can see that the coefficient of $\beta = -0.387 < 0$ for a reference female, main annuitant. This suggests that female main annuitants have a lower hazard compared to female beneficiaries ($\exp\{\beta\} < 1$ times as much)

Summary & Evidence for Chilean Life Tables

In summary, our analysis found statistically significant effects of the covariates of HEALTH and SEX on survival. Particularly, we found that survival is higher for female annuitants compared to male annuitants, and higher for healthy annuitants compared to disabled annuitants. We also found a statistically significant effect of the PERSON_TYPE covariate for female annuitants but not for male annuitants, where survival is (generally) higher for female main annuitants than it is for female beneficiary annuitants. These findings suggest that it is statistically reasonable to separate all annuitants by SEX and HEALTH, and to further separate female annuitants by PERSON_TYPE. These are mostly in accordance with the current Chilean life tables which separate all annuitants by SEX and HEALTH, but only separate healthy (and not disabled) female annuitants by PERSON_TYPE. Investigating further into the disabled female annuitants, we can see that under 800 datapoints were of beneficiaries, whilst almost 40,000 were of main annuitants. This suggests that the mortality estimates for beneficiaries would be highly variable, and thus the set of disabled female annuitants should not be separated by PERSON_TYPE (as done in the current Chilean life tables). Therefore, our statistical analysis supports the use of the current 5 Chilean life tables.

Section 3 – Graduation of Life Table

As established in the survival analysis section above, it was seen that the when the covariate PERSON_TYPE was paired with SEX, it was only statistically significant for female annuitants. Therefore, when required to construct a unisex life table, it is illogical to separate the tables by PERSON_TYPE, and thus we will only graduate one life table for healthy, unisex annuitants. The graduation process only required a small amount of data processing, specifically filtering the dataset by healthy annuitants. The remaining process can be broken down into 3 key stages – calculation of crude estimates, graduating crude estimates and model selection.

Graduation Process & Recommendation

The calculation of crude estimates was a relatively simple process that required a few key assumptions, all of which is explained in Table 6. In short, exposure to risk and deaths at a specific age were calculated using an age last birthday definition, from which mortality was estimated using MLE estimates. The assumed distribution of deaths was a Binomial for q_x estimation, and a Poisson for μ_x. It was also assumed that the estimates applied for the interval $[x, x+1)$. The crude mortality estimates can be found in Table 5.

To graduate the crude estimates, the GM class of models and cubic splines were fitted, the process of which is explained in Explanation 3. To fit a Gompertz and Makeham distribution, the `nls()` function was used on the crude estimates of μ. These were then converted into estimates of q using the calculation specified in Explanation 3. Conversely, 3 splines were fit directly using the estimates of q, with each fit considering a balance between adherence to data and smoothness of graduated curves. This includes a regression spline with knots at $x = 72.5$ and 87.5, and 2 smoothing splines with $spar = 0.63$ and 0.60. The plots for each fit can be found in Explanation 3.

The suggested graduation model was selected based qualitative and quantitative analysis of the fit and smoothness of each curve, which is detailed in Explanation 4. The Gompertz and Makeham models failed the χ^2 test for overall goodness of fit and thus are a poor graduation technique for the given dataset. The smoothing spline with the smaller $spar = 0.6$, which focalised smoothing less in favour of adherence to data, demonstrated relatively large third differences compared to the other remaining models, and thus was suggested. Finally, the remaining smoothing spline demonstrated a decreasing rate of change of mortality for ages above 97, indicating it was capturing the noise at higher ages. Based on our current understanding of mortality and referring to the plot of the current Chilean tables in Figure 14, we are inclined to believe it is capturing noise. Therefore, we recommend the (cubic) regression spline with knots at $x = 72.5$ and 87.5 to graduate unisex Chilean life tables for healthy annuitants. The graduated

Analysis of Graduation Results

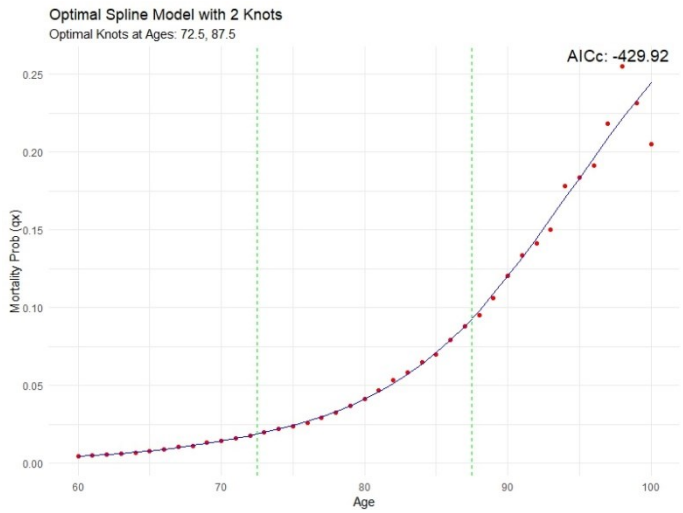

Figure 5 - Suggested Graduation Model

The mortality probabilities shown in Figure 5 showcase a similar relationship to our other graduation techniques. In all our models, see that mortality probability generally increases at an increasing rate. At younger ages closer to 60, relatively low mortality that increases slowly. However, this feature of the data is short lived, with the mortality rate picking up around the early 70s. From our GM class models in Figure 10, we can see that hazard rate deviates from the Gompertz fit. After introducing an additional term to account for this in the Makeham model, we can see that the fit improves significantly, capturing the differing shape at younger ages. A similar effect can be attributed to the knot placed 72.5 years in Figure 5. From here on, all models showcase a much faster increase in mortality probability. However, we see slight differences between all models high ages past 95. Particularly, the GM class models in Figure 10 showcase the mortality rate increasing at an increasing rate, whilst the suggested model states it would increase at roughly a constant rate, and the smoothing splines in Figure 12 suggesting that it would increase at a decreasing rate. However, all models agree that mortality continues to increase.

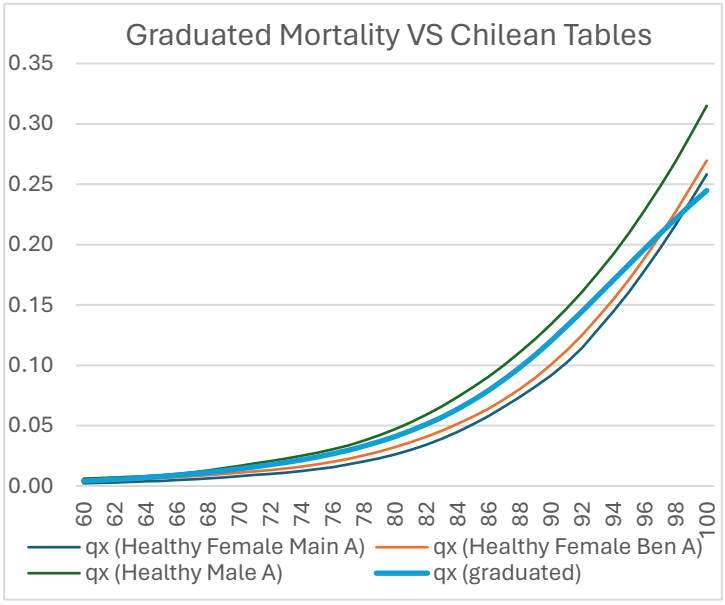

Figure 6 - Graduated Mortality VS Chilean Tables

Compared to the sex specific Chilean life tables for healthy annuitants, we can observe some key characteristics Figure 6. For a start, the mortality probability at an age x in our unisex table is generally between the sex specific mortality. Specifically, the unisex mortality probability is higher than that of female annuitants, but lower than that of male annuitants. Based on our previous analysis, we know that mortality is higher for male annuitants than females. Therefore, by pooling both genders together, we are roughly "averaging out" the effects and thus producing a mortality curve that is not as high as the riskier group (men) but also not as low as the less risky group (females). We can also see that the gradient of our suggested graduated mortality curve is lower for ages beyond the mid-80s. Consequentially, the graduated curve intersects the curves for female annuitants in the Chilean tables. This could be attributed to our suggested regression spline capturing the variance in mortality estimates at older ages.

Section 4 – Ethical Considerations

Using the gender of an annuitant as a rating criteria for pricing annuities has various advantages and disadvantages. Currently, Chilean insurers are allowed to use gender in pricing policies, something which is often considered an ethical grey area. The following section attempts to analyse the sociocultural implications surrounding the use of gender as a rating metric. This will be done by considering the ethical theories of utilitarianism and deontology, and by following a Dobrin Ethical Framework (2008).

In this scenario, the key stakeholders are annuitants (both current and potential), and insurers. The annuitants are interested in receiving a certain level of financial security and predictability by the insurance product, whilst simultaneously assuming its fair and just pricing. The insurers, however, are interested in maximising profitability whilst simultaneously managing financial risk. and further amplify existing social inequalities. This can also lead to the disparate treatment of individuals based on factors they cannot control (their gender), potentially having disproportional impacts of specific demographics within the population. Therefore, the core ethical values that will guide our examination are ensuring the fairness and wellbeing of all stakeholders involved.

A potential course of action is to continue using gender as a rating criteria for pricing annuities. As we have seen in our survival analysis above, there is strong statistical evidence that the mortality of men is higher than that of women. By considering the annuitant's gender, the insurer can set prices that are actuarially sound and more accurately reflect the risk associated with a specific policyholder. It also means that insurers can improve the affordability of their insurance products for subsets of the population which could potentially lead to an increase in profitability. This aligns with the interests of the insurer (effective risk management and profitability), whilst also upholding one of the core values associated with this dilemma (ensuring their financial wellbeing). However, the use of gender as a rating criteria can have negative implications on annuitants. Particularly, gender can capture the effect of other factors such as alcohol abuse, smoking and dangerous driving. These are all factors that occur at higher rates for men and are partly a reasons for the increased mortality in men. This is a concern, as it means that considering gender prices all men according to factors that may not necessarily affect them. This inaccurately makes insurance less affordable and often unavailable to certain groups such as those in a lower socio-economic class. This contradicts our core values of fairness and wellbeing for certain annuitants.

An alternative action is to prevent the use of gender as a rating criteria for pricing annuities. However, this also has some relatively significant consequences on the stakeholders. The removal of gender as a rating factor redistributes the risk associated with higher risk subgroups (men) to those with a lower risk (women). For example, it is well documented that due to their hormonal structure, women have a lower vulnerability towards certain diseases such as heart disease. Considering Chile's pension system, unisex pricing would reduce the retirement income for female annuitants due to the redistribution of risk from males (which can be visualised in Figure 14). This is against the interests of annuitants and violates our core values of fairness (and potential wellbeing) for those affected. From the insurer's perspective, the prevention of gender-based pricing can result in a reduction the accuracy of pricing estimates and increase the risk of insurance provisions. There are various techniques for insurers to deal with this, however a lot of these solutions would require the insurers to investment into research and technology. For example, insurers could invest in technology that provides, direct, specific, and meaningful data of annuitants (such as their daily step count). They could also take an alternative approach to their data analytics and modelling, such as increasing the weight assigned with other rating factors (such as alcohol and smoking habits or health practices).This once again contradicts the interests of the insurer, potentially increasing their risk or reducing their profitability in the short term.

Considering the discussion above through the ethical frameworks of utilitarianism and deontology, it becomes apparent that the practice of using gender as a rating criteria for pricing annuities raises significant ethical concerns. Since using gender as a rating criteria can negatively impact societal welfare and exacerbate inequalities, a utilitarian approach tells us this practice does not maximise overall utility and thus should not be used. Further, since using gender as a rating criteria violates key principles of deontology by contributing towards discrimination, unfairness, and inequality, we are further convinced to reject this approach. Instead, we recommend insurers replace gender with the potential underlying factors that directly contribute to the differences in mortality. This approach benefits most of the population (maximising utility) and upholds key ethical principles (thus being deontologically acceptable).

Section 6 – Appendix

Appendix 6.1 – Data Preparation & Cleaning

In preparation for the descriptive analysis, the mortality data was cleaned and processed. It was found that whilst there were no NA values in the data, there were 69 rows in which `DATE_START = DATE_END`. This was an anomaly in the data as it did not contribute towards exposure and produced NA values during the survival analysis stage. This represents less than 0.006% of the dataset, thus the impact of these rows was deemed insignificant, and they were removed from the dataset. Further, the mortality dataset contained observations of people who were less than age 60 on January 1st, 2014. Whilst these observations might seem like anomalies, these were all people who turned 60 before the end of the investigation, and thus these data points kept in the investigation and used to gain further insights in the survival analysis section.

Explanation 1 - Data Preparation & Cleaning Process

Appendix 6.2 – Descriptive Analysis Statistics & Plots

Age Groups	Death Proportion
60-65	0.0481
65-70	0.0520
70-75	0.0724
75-80	0.104
80-85	0.170
85-90	0.233
90-95	0.296
95+	0.371

Table 1 – Death Proportion by Age Group

Covariate	Value	Death Proportion
Sex	Male	0.107
Sex	Female	0.0601
Health	Disabled	0.148
Health	Healthy	0.0739
Person Type	Beneficiary	0.0625
Person Type	Main Annuitant	0.0943

Table 2 – Death Proportion by Covariates

	Sex	Health	Person Type
Sex	1.000	-0.140	0.579
Health	-0.140	1.000	-0.267
Person Type	0.579	-0.267	1.000

Table 3 - Correlation Matrix of Covariates

Appendix 6.3 – Cox Proportional Hazard Regression Model (with Interaction)

```
> summary(best_model_with_interaction)
Call:
coxph(formula = Surv(time = mortality$age_at_start, time2 = mortality$age_at_end,
    event = mortality$DEATH) ~ HEALTH + SEX + PERSON_TYPE:SEX +
    PERSON_TYPE:HEALTH, data = mortality, method = "breslow")

  n= 1291958, number of events= 103827

                                          coef exp(coef) se(coef)       z Pr(>|z|)
HEALTHHealthy                          -1.21944   0.29539  0.08047 -15.153  < 2e-16 ***
SEXM                                    0.23118   1.26009  0.02081  11.109  < 2e-16 ***
SEXF:PERSON_TYPEMain Annuitant         -0.38748   0.67877  0.08126  -4.769 1.85e-06 ***
SEXM:PERSON_TYPEMain Annuitant          0.03369   1.03426  0.08095   0.416   0.6773
HEALTHHealthy:PERSON_TYPEMain Annuitant 0.13348   1.14280  0.08095   1.649   0.0992 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                       exp(coef) exp(-coef) lower .95 upper .95
HEALTHHealthy                             0.2954     3.3853    0.2523    0.3459
SEXM                                      1.2601     0.7936    1.2097    1.3125
SEXF:PERSON_TYPEMain Annuitant            0.6788     1.4733    0.5788    0.7960
SEXM:PERSON_TYPEMain Annuitant            1.0343     0.9669    0.8825    1.2121
HEALTHHealthy:PERSON_TYPEMain Annuitant   1.1428     0.8750    0.9751    1.3393

Concordance= 0.624  (se = 0.001 )
Likelihood ratio test= 20735  on 5 df,   p=<2e-16
Wald test            = 24947  on 5 df,   p=<2e-16
Score (logrank) test = 27334  on 5 df,   p=<2e-16
```

Figure 7 - Cox Regression Model with Interaction

Appendix 6.4 – Proportional Hazard Assumption for Cox Regression Model

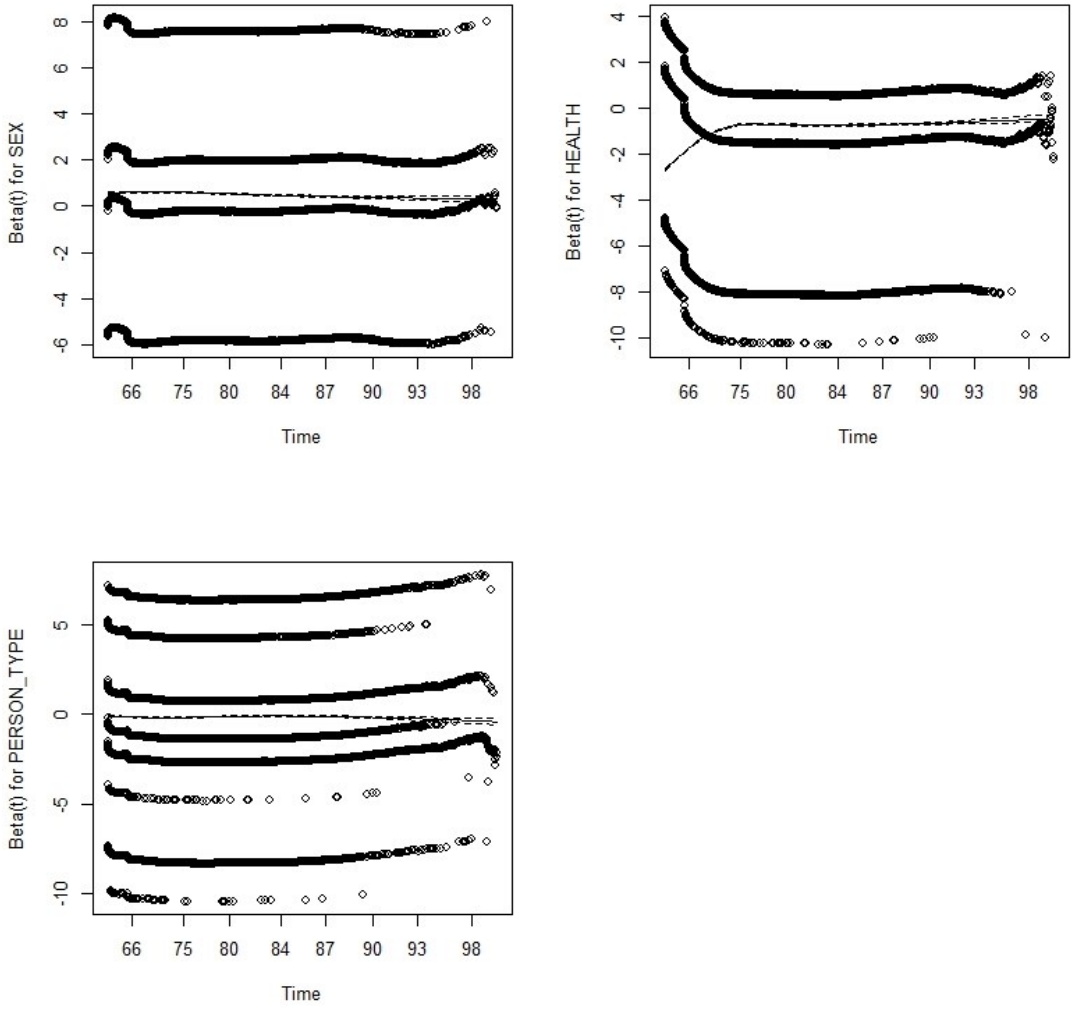

Figure 8 - Schoenfeld Residuals for Cox Regression Model (without Interaction)

Covariate	Shoenfeld Individual Test p-value
SEX	$< 2 \times 10^{-16}$
HEALTH	$< 2 \times 10^{-16}$
PERSON_TYPE	$< 2 \times 10^{-16}$

Table 4 - Schoenfeld Individual Test p-values for Cox Regression Model (without Interaction)

The Cox regression model relies on the proportional hazards assumption, which states that there is effect of a covariate on the hazard function is independent of time. This means that in theory, Schoenfeld residuals should be independent of time. Thus, a plot that shows a non-random pattern against time is evidence of a violation of the proportional hazards assumption. When the Schoenfeld residuals are plotted for the 3 covariates SEX, PERSON_TYPE and HEALTH in Figure 8, we can see a clear relationship between residuals and time. Further, performing the Schoenfeld individual test, we get a p-value less than 2×10^{-16} for each covariate. This means we are inclined to reject the null hypothesis of the test, which states that Schoenfeld residuals are independent of time. Therefore our statistical analysis indicates that the Cox regression model violates the proportional hazards assumption.

Explanation 2 - Justification of Violation of Proportional Hazards Assumption

Appendix 6.5 – Crude Estimates

x	E_x (last)	E_x^C (last)	d_x (last)	$\widehat{q}_x$	$\widehat{\mu}_{x+0.5}$
60	166202.5	165872.4312	690	0.004152	0.0041598
61	193411.3	192957.165	930	0.004808	0.0048197
62	212822	212250.0876	1171	0.005502	0.0055171
63	227514.4	226848.0205	1351	0.005938	0.0059732
64	238751.5	237974.165	1632	0.006836	0.0068579
65	308676.8	307542.2649	2390	0.007743	0.0077713
66	315807.2	314388.4264	2867	0.009078	0.0091193
67	310590.1	309043.193	3171	0.01021	0.0102995
68	299513.1	297852.4244	3379	0.011282	0.0113445
69	284559	282728.193	3740	0.013143	0.0132283
70	266887.4	264944.4702	3885	0.014557	0.0146635
71	246532.4	244610.7974	3909	0.015856	0.0160663
72	225238.4	223247.9528	3981	0.017675	0.0178322
73	204206.7	202201.4052	3991	0.019544	0.0197377
74	183254.4	181240.5195	4016	0.021915	0.0221584
75	161413.3	159495.7132	3803	0.023561	0.0239003
76	141887.8	140029.5175	3695	0.026042	0.0263873
77	124048	122214.8713	3644	0.029376	0.0298163
78	107537.5	105762.6324	3505	0.032593	0.0331402
79	92151.48	90433.44969	3413	0.037037	0.0378621
80	79395.97	77784.35729	3256	0.04101	0.0418593
81	68440.56	66830.24914	3195	0.046683	0.0478077
82	58489.88	56896.51061	3114	0.05324	0.0547309
83	49652.69	48173.83299	2888	0.058164	0.0600949
84	41557.88	40198.31828	2693	0.064801	0.0669929
85	34240	33021.86516	2396	0.069977	0.072558
86	27518.31	26396.58658	2185	0.079402	0.0827759
87	21216.37	20252.09788	1862	0.087762	0.0920892
88	15679.82	14904.07734	1492	0.095154	0.1001068
89	11431.19	10802.35113	1211	0.105938	0.1121052
90	8061.88	7567.03833	972	0.120567	0.1284518
91	5498.99	5114.03217	734	0.133479	0.1437222
92	3795.855	3520.930869	536	0.141207	0.1522325
93	2622.489	2406.951403	394	0.150239	0.1636925
94	1718.604	1559.1013	306	0.178051	0.1962669
95	1109.375	1007.299795	204	0.183887	0.2045071
96	694.3552	626.2340862	133	0.191545	0.2123806
97	407.3333	358.6173854	89	0.218494	0.2481754
98	235.3018	203.9123888	60	0.254992	0.294244
99	138.0808	119.7508556	32	0.231748	0.2672215
100	87.85421	78.82888433	18	0.204885	0.2283427

Table 5 - Crude Mortality Estimates

Appendix 6.6 – Crude Estimate Calculation & Assumptions

Variable	Calculation	Assumptions
E_x^C	If an individual i is exposed to risk between an age $[x, x+1)$, then their central exposure to risk for an age x is: $$E_{x,i}^C = b_{x,i} - a_{x,i}$$ Where: $$b_{x,i} = \min(\text{age at exit}_i, x+1)$$ $$a_{x,i} = \max(\text{age at entry}_i, x)$$ Otherwise, $E_{x,i}^C = 0$. $$E_x^C = \sum_{\text{all } i} E_{x,i}^C$$	Using an age last birthday definition. Exposure starts at the start of the day of entry and ends at the start of the day of exit.
E_x	If the individual i is exposed to risk between an age $[x, x+1)$, then their initial exposure to risk for an age x is: $$E_{x,i}^c = b_{x,i} - a_{x,i}$$ Where: $$b_{x,i} = \begin{cases} x+1 \text{ if they die in } [x, x+1) \\ \min(\text{age at exit}_i, x+1) \text{ otherwise} \end{cases}$$ $$a_{x,i} = \max(\text{age at entry}_i, x)$$ Otherwise, $E_{x,i} = 0$ $$E_x = \sum_{\text{all } i} E_{x,i}$$	Using an age last birthday definition. Exposure starts at the start of the day of entry and ends at the start of the day of exit.
d_x	Number of observations where individual dies and $\lfloor \text{age at death} \rfloor = x$	Using an age last birthday definition.
$\hat{q}_x$	$$\hat{q}_x = \frac{d_x}{E_x}$$	IID observations. $\hat{q}$ is an estimate derived from Binomial MLE. Using an age last birthday definition. $\hat{q}_x$ is the estimate at the age at the start of the rate interval and applies for $[x, x+1)$.
$\hat{\mu}_{x+0.5}$	$$\hat{\mu}_{x+0.5} = \frac{d_x}{E_x^C}$$	IID observations. $\hat{\mu}$ is an estimate derived from Poisson MLE. Using an age last birthday definition. $\hat{\mu}_{x+0.5}$ is the estimate at the middle of the rate interval and applied for $[x, x+1)$

Table 6 - Crude Estimate Calculations & Assumptions

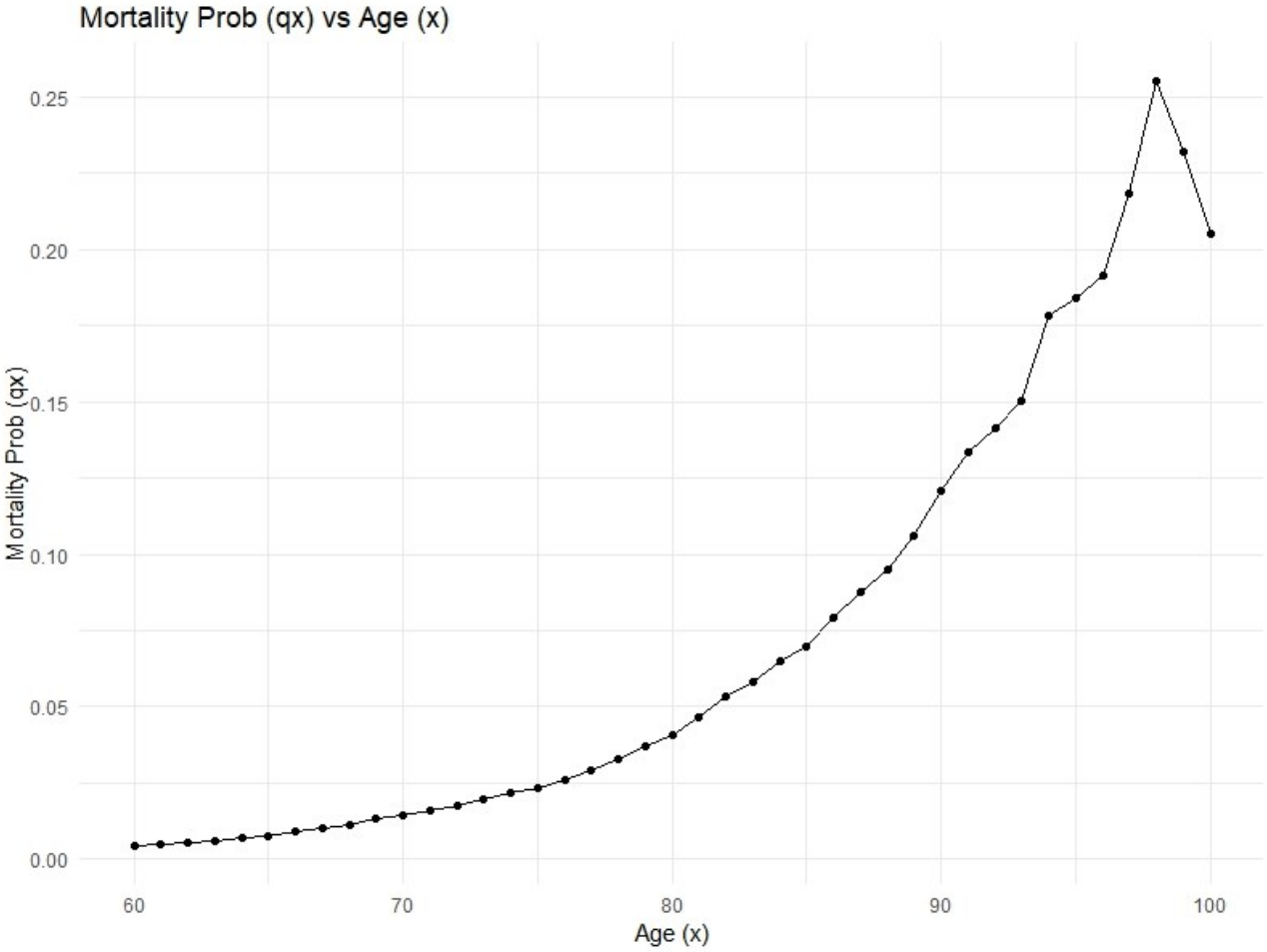

Figure 9 - Crude Estimate Plot (q_x)

Appendix 6.7 – Pre-selected Graduation Models

Explanation 3 - Graduation Model Pre-Selection & Graduation Models

The Gompertz and Makeham models were fit using the crude estimates $\hat{\mu}_{x+0.5}$ found in Table 6. This gives a continues graduated curve for each model, $\dot{\mu}_t$. Then, these graduated rates were converted to mortality probabilities $\dot{q}_x$ using the following equation below. This assumes that the probability of death is equal throughout the interval $[x, x+1)$.

$$\dot{q}_x = \int_x^{x+1} \dot{\mu}_t \, dt.$$

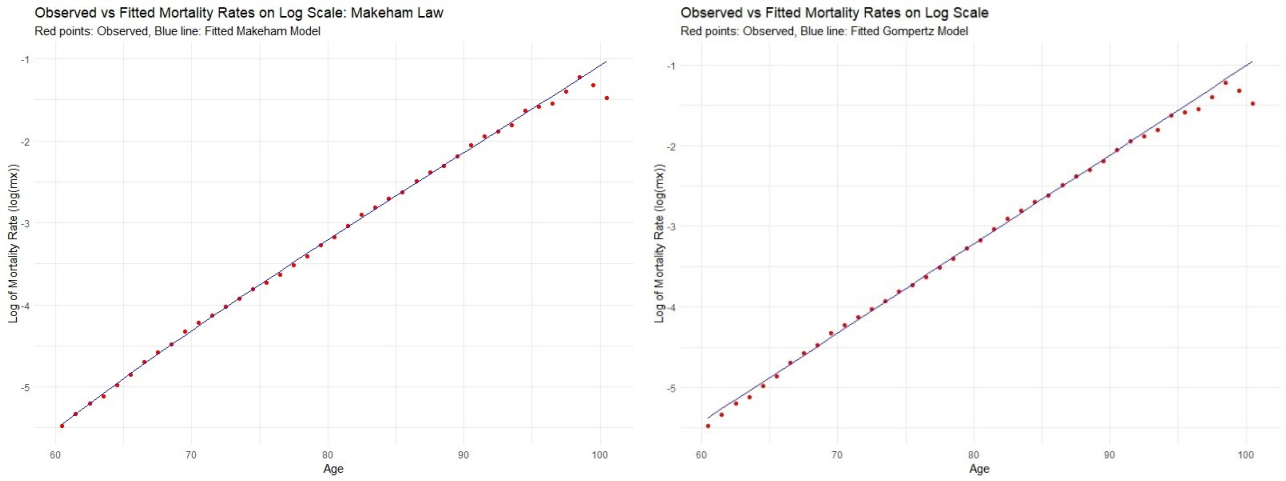

Figure 10 - GM Model Plots (log)

Based on our survival analysis, it could be seen that there were distinct features in the mortality distribution of our crude estimates. Particularly, that mortality increases at an increasing rate with age. Further, looking at the plot of our crude estimates in Figure 9, it seems reasonable to place 2 knots, one to capture younger ages (roughly between 60 and 75), and one to capture older ages. An algorithm was developed to fit a regression spline to the crude estimates for $\hat{q}_x$ in Table 6. A restriction was set so that the algorithm could not place a knot between age $[95, 100]$ to avoid overfitting to the noise present at older ages. Further, a weight $w_x = E_x \div \hat{q}_x$ was used to further minimise overfitting to older (as they have a lower exposure and higher estimated mortality probability. This algorithm subdivided the x-axis into 50 equidistant points and attempted to place k (specified) knots on a set of k of these points. Tracked and selected the best knot locations based on AIC. This algorithm was run for $k = 2$ and $k = 3$, which were decided by observation (2 knots to account for the distribution of older and younger annuitants, and perhaps a third knot in between to account for the distribution of other annuitants). It was noticed that the algorithm consistently placed the first two knots very close to each other (within 5 years, generally around 65 and 70). This seemed unnecessary due to the similarity in distribution of such ages, therefore the following regression spline with 2 knots was put forward.

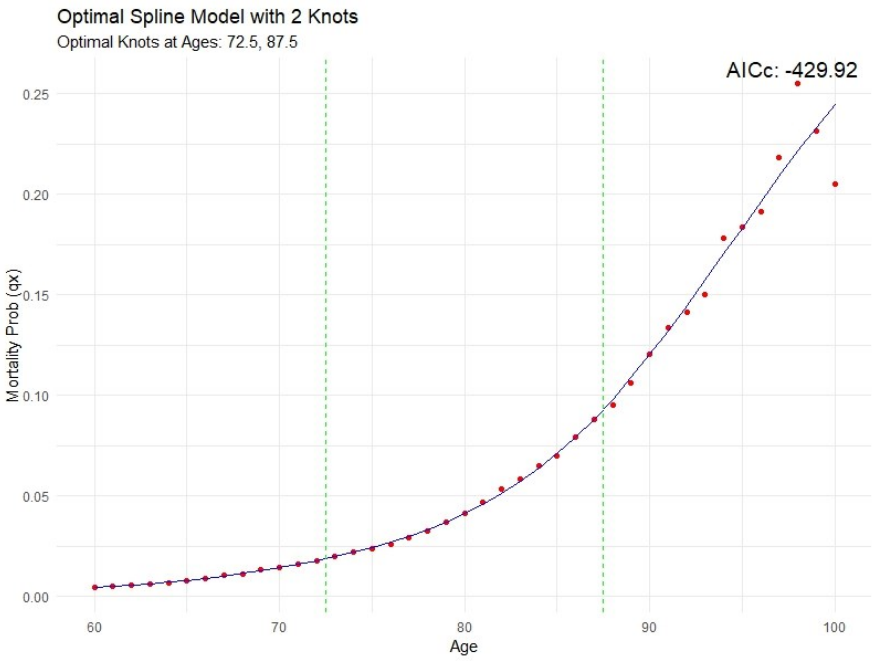

Figure 11 - Optimal Regression Spline Plot

This regression spline was set as the standard for the quality of fit for smoothing splines. An algorithm was developed that tests all possible *spar* values between 0.01 and 0.99 by indexes of 0.01 and prints all *spar* values that fit a smoothing spline with a lower *RSS* than the regression spline in Figure 11. Then, 2 smoothing splines were selected based on trial and error, one with $spar = 0.60$ and another with $spar = 0.63$ to represent a models emphasises adherence to data more and smoothness slightly more (respectively).

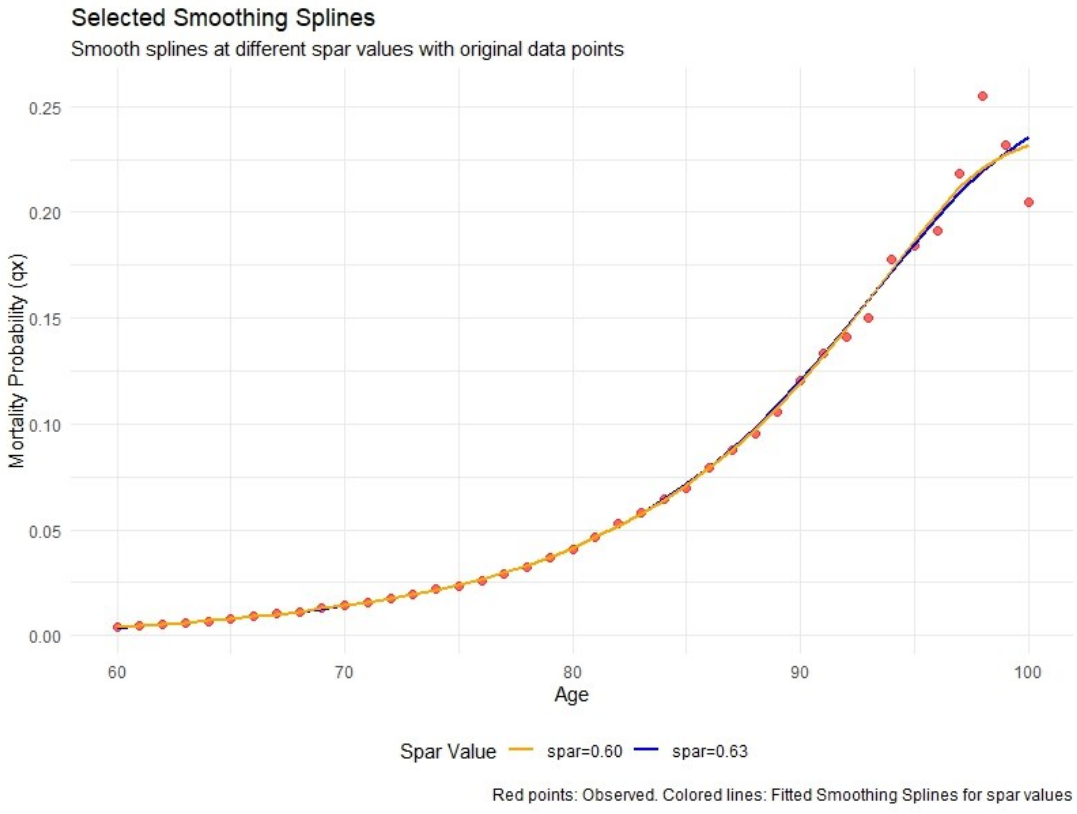

Figure 12 - Selected Smoothing Spline Plots

Appendix 6.8 – Statistical Tests & Smoothness Tests

Model	χ^2	Signs Test	Cum. Dev. Test	Grouping of Signs Test	Std Deviance Test
Gompertz	1.532×10^{-5}	0.349	0.969	8.020×10^{-5}	8.705×10^{-6}
Makeham	0.032	0.755	0.846	0.024	0.375
Regression Spline ($k = 2$)	0.641	1.000	0.916	0.216	0.989
Smoothing Spline ($Spar = 0.60$)	0.739	0.349	0.736	0.252	0.970
Smoothing Spline ($Spar = 0.63$)	0.147	0.755	0.442	0.081	0.847

Table 7 - Statistical Test Results (p-values)

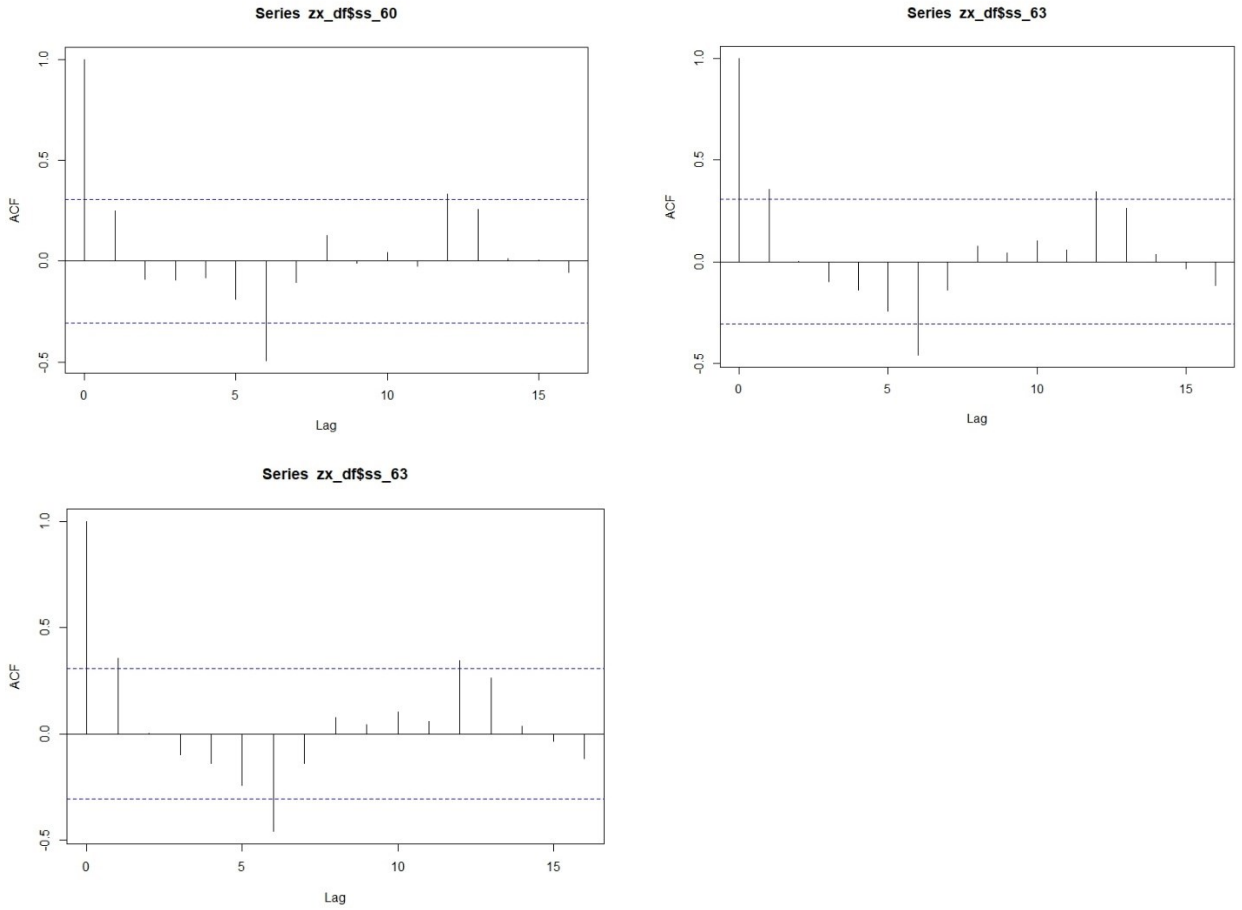

Figure 13 - Serial Correlations of Spline Models

x	Regression Spline	Smoothing Spline (Spar $= 0.63$)	Smoothing Spline (Spar $= 0.60$)
63	6.039E-06	1.876E-05	2.615E-05
64	6.039E-06	2.225E-05	3.124E-05
65	6.039E-06	1.791E-05	2.135E-05
66	6.039E-06	1.175E-05	9.003E-06
67	6.039E-06	6.114E-06	-1.901E-06
68	6.039E-06	5.821E-06	-1.650E-06
69	6.039E-06	6.246E-06	-1.748E-06
70	6.039E-06	6.332E-06	-5.013E-06
71	6.039E-06	1.566E-05	1.073E-05
72	6.039E-06	2.591E-05	2.713E-05
73	6.964E-06	3.174E-05	3.305E-05
74	2.825E-05	3.838E-05	4.156E-05
75	4.953E-05	4.642E-05	5.504E-05
76	5.045E-05	5.590E-05	7.433E-05
77	5.045E-05	5.196E-05	6.539E-05
78	5.045E-05	4.182E-05	4.396E-05
79	5.045E-05	3.439E-05	2.879E-05
80	5.045E-05	2.555E-05	8.564E-06
81	5.045E-05	2.295E-05	-6.243E-07
82	5.045E-05	1.806E-05	-1.874E-05
83	5.045E-05	3.075E-05	9.642E-07
84	5.045E-05	6.330E-05	6.714E-05
85	5.045E-05	8.170E-05	1.066E-04
86	5.045E-05	8.636E-05	1.251E-04
87	5.045E-05	5.824E-05	8.223E-05
88	4.669E-05	4.350E-05	8.524E-05
89	-3.988E-05	1.405E-05	6.879E-05
90	-1.264E-04	-6.202E-05	-4.381E-05
91	-1.302E-04	-1.324E-04	-1.388E-04
92	-1.302E-04	-1.402E-04	-9.329E-05
93	-1.302E-04	-1.436E-04	-4.887E-05
94	-1.302E-04	-2.575E-04	-2.806E-04
95	-1.302E-04	-4.049E-04	-6.274E-04
96	-1.302E-04	-3.340E-04	-5.237E-04
97	-1.302E-04	-3.542E-04	-6.813E-04
98	-1.302E-04	-4.275E-04	-1.000E-03
99	-1.302E-04	-1.121E-04	-4.300E-04
100	-1.302E-04	6.188E-04	1.157E-03

Table 8 - Third Differences for Splines

Appendix 6.9 – Model Selection

Explanation 4 - Model Selection Logic

Based on qualitative assessment, we can note a few things. Firstly, from Figure 10, we can see that the Gompertz model fails to accurately model the mortality probability for ages closer to 60. This is not the case for the Makeham model, which contains an additional adjustment term A which considers this effect. Next, we can see from Figure 12 that the two smoothing splines are highly similar from ages 60 to 95. However, the graduated curves slightly deviate for the last few (highly variable) data points. We observe that both graduated curves demonstrate that the mortality probability increases at a slower rate between ages 97-100. This suggests that the smoothing splines may be capturing the variance at older ages, an issue that is much more prevalent in the spline with $spar = 0.60$. This could mean an overemphasis towards adherence to data, as we generally would not expect mortality to slow down past 95. Thus, we are inclined to believe that the Gompertz model is a poor fit overall, and the smoothing splines may be capturing the noise in the data for older ages.

Looking at the statistical tests in Table 7, we can start by considering the χ^2 overall GOF test. This tests for statistical discrepancies between the crude estimates and graduated rates. As the p-value is less than 0.05 for the Gompertz and Makeham fits, we are inclined reject the null hypothesis at a level $\alpha = 0.05$. This means that the deviations in graduated estimates are too large (over-graduation). This aligns with our qualitative examination for the Gomperz model. The other 3 spline uphold a relatively strong overall goodness of fit, and thus from here we will only consider the 3 spline models for our recommended method for graduation.

Next, we aim to assess the direction of bias. Looking at the Signs Test results in Table 7, we cannot see anything indicating that any models demonstrate an unequal balance of positive and negative deviations. Whilst this indicates that there isn't a direction bias in general, it does not examine the bias in specific regions within the data. To do this, we consider the Grouping of Signs Test results in Table 7. Once again, we can see that the p-value is greater than 0.05 for all 3 spline models, meaning we have no statistical evidence that concludes that the rates may be consistently too high or too low over certain parts. This is further supported by Figure 13, which showcases that the deviations are independent at consecutive ages (lack of serial correlation).

Then, looking at Cumulative Deviations test, we can test whether the number of deaths conforms to the mortality rates assumed in graduation. Once again, each of the spline models demonstrate a p-value above 0.05, suggesting that a level $\alpha = 0.05$, we have insufficient statistical evidence to conclude that the actual variance is higher than predicted by the assumed model for the range of ages considered.

Finally, considering the results in Table 8, we observe third differences of a magnitude smaller than 10^{-4} for the smoothing spline with $spar = 0.6$ at ages beyond 97. This is relatively small compared to the other two splines, which all showcase third differences of magnitudes higher than 10^{-4} for all ages. This aligns with the insights from our qualitative analysis above – that the spline may be capturing the noise in the data at older ages. Therefore, from here, we put forward the other two models, the regression spline and smoothing spline with $spar = 0.63$.

From here, we can consider the remaining 2 models from a qualitative and algorithmic perspective. Since we know that the smoothing spline was specifically selected such that it would have a lower RSS than the regression spline, the question arises whether this smoothing spline is capturing any additional noise in the data that may not be picked up by the tests discussed above. We also know that the algorithm for the regression spline was specifically designed to avoid capturing this, therefore it is less likely to be an issue here as supported by Figure 11. Finally, Figure 14 from the current Chilean life tables does not showcase a decrease in the rate of mortality risk at ages beyond 95. Therefore, based on the qualitative suspicions discussed above, we select the regression spline with 2 knots as the model to graduate our mortality probabilities.

Appendix 6.10 – Recommended Graduation Model & Life Table:

x	$\dot{q}_x$
60	0.00415
61	0.00472887
62	0.00538806
63	0.0061336
64	0.00697153
65	0.0079079
66	0.00894874
67	0.01010009
68	0.01136799
69	0.01275847
70	0.01427759
71	0.01593136
72	0.01772584
73	0.01966799
74	0.02178606
75	0.02412956
76	0.02674897
77	0.02969473
78	0.0330173
79	0.03676713
80	0.04099468
81	0.0457504
82	0.05108474
83	0.05704817
84	0.06369114
85	0.07106409
86	0.07921749
87	0.08820179
88	0.09806368
89	0.10876328
90	0.12017414
91	0.13216606
92	0.14460882
93	0.1573722
94	0.170326
95	0.18334001
96	0.19628401
97	0.20902779
98	0.22144114
99	0.23339384
100	0.24475569

Table 9 - Recommended Graduated Life Table

Appendix 6.11 – Other Supporting Plots

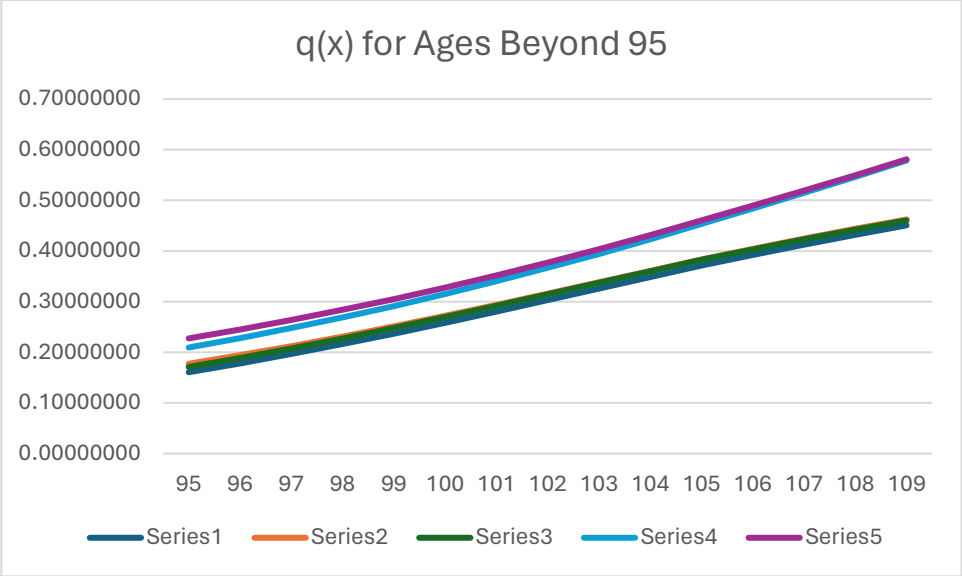

Figure 14 - Mortality Probability for Ages Beyond 95

Section 7 – Use of AI

Use of AI throughout this assignment consisted strictly of ChatGPT. AI was not used in the writing of this report for tasks 1 to 3; however, it was used to generate some R code. Specifically, it was used to generate the code for plots and improve the aesthetic of all plots presented. Beyond this, it was also used to generate repetitive code for copying data into new data frames. For task 4, AI was used to assist in researching and understanding the ethical implications of pricing insurance products using gender.

Examples of some of the prompts used are listed below:

- "Here is my code for plotting X, Y and Z. Send me the R code to improve the aesthetic of these plots. Ensure these plots are appropriate for a professional report."

- "Here is my code this far {code}. Write me an R script that plots a X plot of Y variable. Ensure these plots are appropriate for a professional report".

- "Here is my code so far {code}. Repeat this for the following X data frames {code}.

- "List me some reasons against gender-based insurance pricing.

To the best of my knowledge, all the other code in this assignment was written using previous understanding, RTutorials or the references on the next page.

Section 8 – References

- ACTL3141 Slides (mod 2-7)

- ACTL3141 RTutorials (Survival Analysis, Graduation)

- https://dr.lib.iastate.edu/server/api/core/bitstreams/40ec8d92-910e-4d15-8032-2e523285f7ef/content#:~:text=Schoenfeld%20residuals%20are%20intended%20to,form%20for%20the%20observed%20responses.

- https://rpubs.com/kaz_yos/resid_co

- https://dr.lib.iastate.edu/server/api/core/bitstreams/40ec8d92-910e-4d15-8032-2e523285f7ef/content#:~:text=Schoenfeld%20residuals%20are%20intended%20to,form%20for%20the%20observed%20responses