

# Improving Home Insurance Affordability

## Adding Bushfire Risk to the ARPC Cyclone Reinsurance Pool

William Keir

### 1. Introduction and Background

In recent years, Australia has been faced with a growing challenge to make home insurance affordable. The increasing risks of natural catastrophes, including cyclones, bushfires, floods and storms, has resulted a significant rise in the frequency and severity of insurance claims, influencing an upwards pressure on home insurance premiums, particularly in areas more prone to such catastrophes. Many households are beginning to struggle to afford home insurance, causing customers to drop out of their policies as they can't afford the coverage they need. This poses a significant risk to homeowners and insurers alike, as homeowners are missing out on having their risks covered, while insurers lose business, creating insolvency pressures on companies.

To tackle this rising issue, our group, acting as an actuarial advisory group, proposes to add bushfire risk to the pre-existing Cyclone Reinsurance Pool (CRP) offered by the Australian Reinsurance Pool Corporation (ARPC). The aim of this proposal is to further diversify the risk pooled within the CRP, allowing for a reduction in the reinsurance premium rates charged to insurers, allowing them to pass savings down to the customer, ultimately reducing affordability pressures on home insurance. Insurers benefit from a more secure, government-backed reinsurance pool with cheaper premiums, reducing their insolvency pressures and enabling them to improve customer confidence. Homeowners benefit from a more secure insurer with reduced premiums, influencing them to hold onto their policies and the benefits that come with it.

### About The ARPC CRP

The ARPC's CRP is a government-backed program designed to provide reinsurance coverage for insurers against losses incurred from claims relating to cyclones and other related natural disasters. It aims to ensure that homeowners in high-risk cyclone areas can access affordable insurance coverage. The CRP effectively pools risks across participating insurers, spreading the financial risk associated with cyclones and related disasters.

Adding bushfire risk to the ARPC's CRP further diversifies the risks covered, reducing reinsurance premiums for insurers. Insurers are incentivised to provide coverage in high-risk locations, increasing access to protection against natural catastrophes. As a result, sustainability of the insurance market improves, benefiting both insurers and consumers in the long run.

### Modelling Improvements to Home Insurance Affordability

From an actuarial perspective, implementing this proposal will require robust modelling and risk assessment to ensure that home insurance affordability improves. Models that accurately map current pricing data will need to be constructed to assess the current affordability issues. After the strategy has been implemented, new data on average premium rates should reflect any improvement that resulted in the changed structure of the CRP. A new model can be fitted to this new data, with comparisons made to the old model. Improvements in home insurance affordability should be reflected in the differences between the old and new models. From there, we can determine whether the strategy worked, and whether more changes need to be made in order to make further improvements.

## 2. Preliminary Data Analysis

### 2. a. Data Cleaning

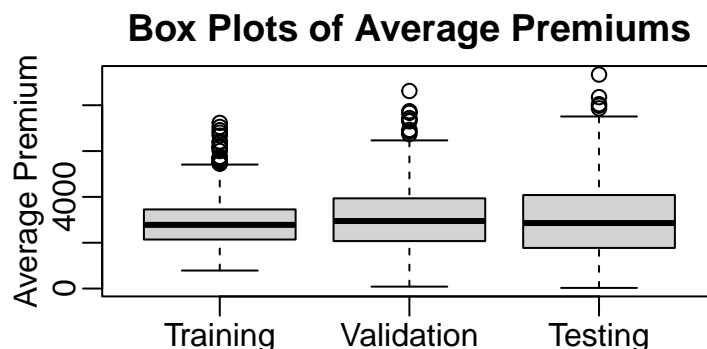
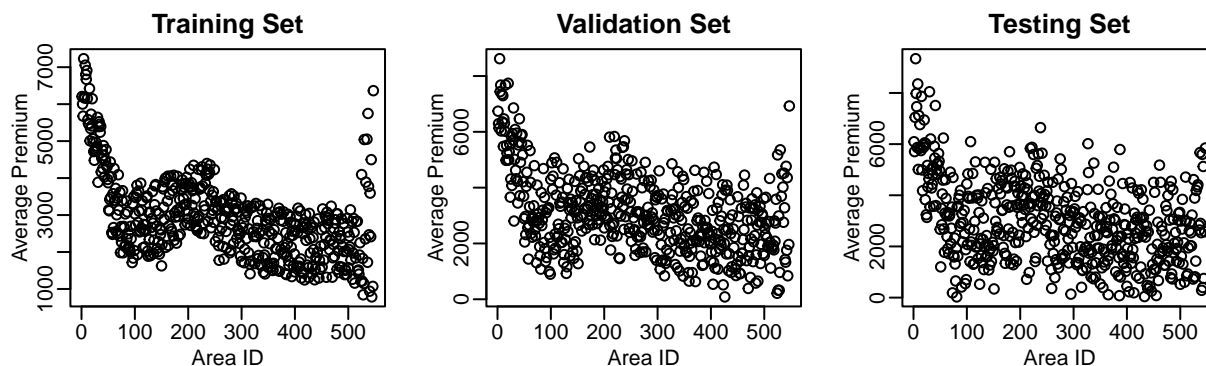
First thing to note is the `Average.cost.of.4.weeks.premium` column appears to be of character format. Also, the `Validation.Data` and `Test.Data` columns contain negative values. In this context, negative values don't make sense, as insurance premiums are always positive values. Further, 2 extremely small outliers exists in the `Average.cost.of.4.weeks.premium` column. Keeping these values in will only hurt the models' ability to capture the overall trend.

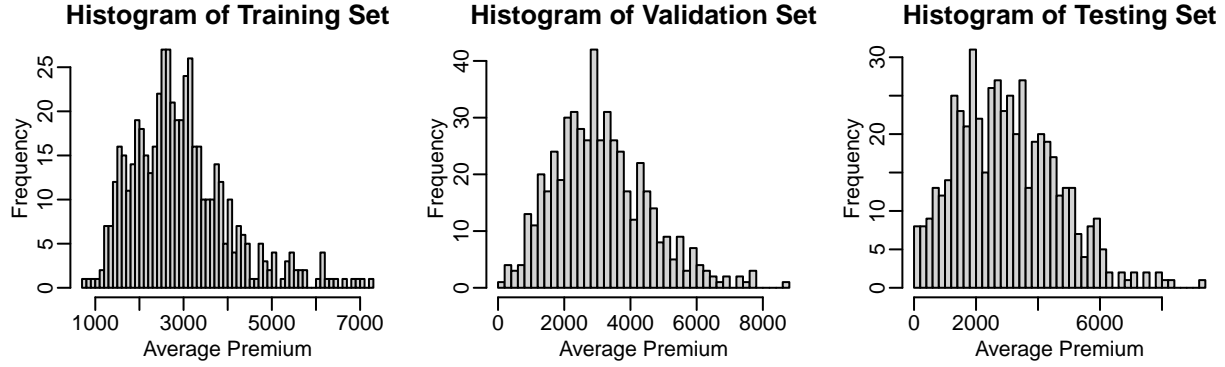
To solve these issues, the `Average.cost.of.4.weeks.premium` column was changed to numeric format. Doing so replaces any character values with NA's. Upon doing so, 3 NA's were produced. For any row containing NA's, the entire row was removed. Rows containing any negative values were also removed. For the two extremely small values were found in the `Average.cost.of.4.weeks.premium` column, both rows containing these values were removed. Removing entire rows in this manner helps maintain consistency and accuracy within the dataset, reducing the possible impact of biases, and helping our code execute seamlessly without having to make adjustments for missing values.

### 2. b., c. Data Exploration and Explanations

Reviewing the data after cleaning, and plotting:

##	Area.ID	Average.premium	Validation.Data	Test.Data
##	Min. : 1.0	Min. : 785.6	Min. : 84.25	Min. : 28.09
##	1st Qu.:132.2	1st Qu.:2140.3	1st Qu.:2072.38	1st Qu.:1773.23
##	Median :264.5	Median :2779.3	Median :2947.02	Median :2858.69
##	Mean :268.0	Mean :2932.7	Mean :3099.09	Mean :3003.85
##	3rd Qu.:402.8	3rd Qu.:3452.6	3rd Qu.:3932.47	3rd Qu.:4081.41
##	Max. :547.0	Max. :7229.7	Max. :8621.66	Max. :9338.66





From the histograms, we see the distribution of the average premium data appears to be approximately normal, with a small amount of skewness making the data appear left-skewed.

From the boxplots, the average premium data appears to be similarly distributed, with the same mean. The testing set appears to have slightly more variance than the validation, which also appears to have slightly more variance than the training set.

From the scatterplots, we can see a distinct pattern in the training set, where it appears linear with a steep negative gradient for Area ID values less than 100, into a slight uptrend from 100 to 225, into a convex downtrend from 225 to 500, then for Area ID values greater than 500, the data seems to have extremely high variation, with no clear trend. The larger values for these ID's appear to be very sparse, to the point where they may be considered outliers, but since there are quite a few points here, we will not remove them.

Such variation in the data suggests a relatively complex model will be needed to fit this data, in order to effectively capture the changing trends. Some examples of such models may include polynomial regression models of degree 4 or greater, or splines.

Also seen in the scatterplots, the validation set appears to possess the same trend as the training set, with greater variation. The testing set has even more variation, although a rough trend can be made out from it.

High variation and few datapoints beyond 500 implies the model will easily overfit to the data if it is too flexible. Large shifts in the trends within the data, such as around ID = 100 and ID = 225, suggests there may be some underlying factors of homes within these Area ID's that are impacting their home insurance premiums, for example, a strong shift in relative location, socio-economic and demographic factors, age and condition of the homes etc.

The whole dataset appears to undergo changes in variance, with slightly lower variation in ID's below 100, with slightly higher variation between 100 and 500, and extremely high variation in points beyond 500. This suggests that the residuals of any model we fit will be heteroskedastic, which may impact the validity of the assumptions made for testing the model.

In our modification of the ARPC's CRP, adding bushfire risk will indirectly aid home insurers in reducing premiums for their customers, particularly for those paying higher rates in high risk areas. This will be reflected in the data by an overall reduction in premiums, with the largest reductions occurring in areas with the highest premiums. The dataset will appear to have less variation, with smoother trends.

## 2. d. Justification of Graduation

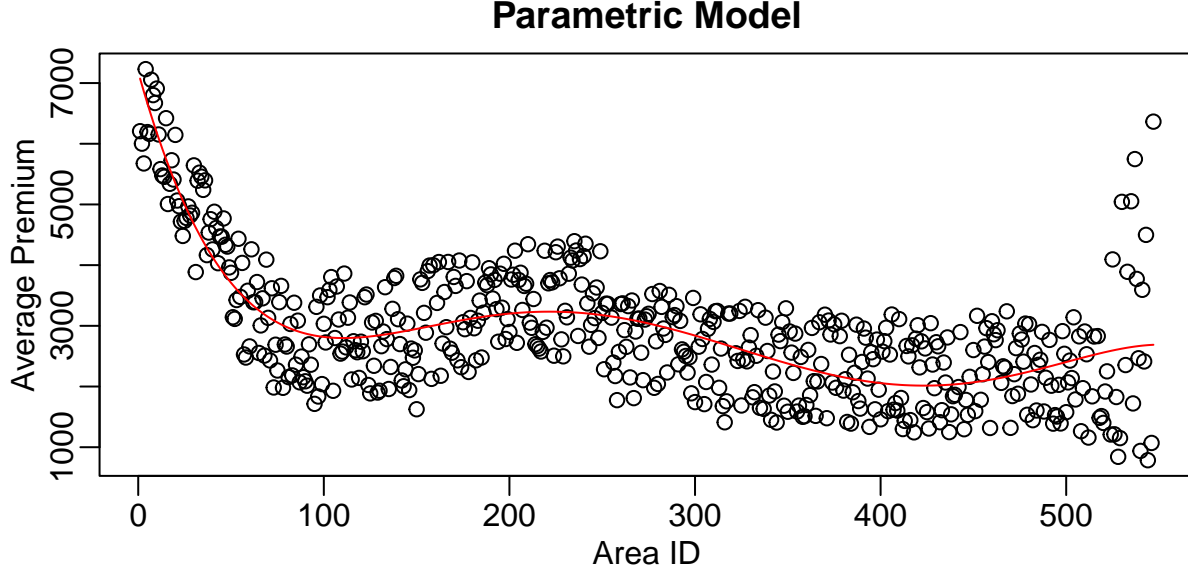
Clearly from the above scatterplots, there are distinct shifts in the trends within the data. Due to noise in the data, it can be difficult to visualise exactly where these shifts in trend occur. Applying graduation techniques help to smooth out variations within the data, making shifts in trend clearer. This allows for a better understanding of the general structure of the data, making it more interpretable.

Due to this, graduated data can be used to make more accurate predictions and forecasts, as it highlights the underlying trends without being skewed by random fluctuations. This forms a strong basis to help insurers in decision making, risk management, resource allocation and planning. This is essential in mapping changes to home insurance premiums, enabling insurers and other entities to address affordability issues for policyholders.

### 3. Parametric Curve Fitting

#### 3. a. Parametric Formula

From the visualisation plots, it is clear to see the data follows what appears to be a ‘W’ shape, with three distinct critical points. From this, we could try fitting a polynomial model of degree 4 or higher to the data. Upon testing, it was found that a polynomial model of degree 5 fits the data better, determined by a combination of visual inspection and significance of coefficients, shown below.



##		Estimate	Std. Error	t value	Pr(> t )
##	(Intercept)	7.182461e+03	1.842327e+02	38.985806	6.507817e-156
##	poly(ID, 5, raw = TRUE)1	-1.094665e+02	6.811756e+00	-16.070231	2.061087e-47
##	poly(ID, 5, raw = TRUE)2	9.705607e-01	7.718916e-02	12.573795	8.133911e-32
##	poly(ID, 5, raw = TRUE)3	-3.711468e-03	3.583272e-04	-10.357763	5.732111e-23
##	poly(ID, 5, raw = TRUE)4	6.254355e-06	7.236666e-07	8.642592	6.931693e-17
##	poly(ID, 5, raw = TRUE)5	-3.828787e-09	5.276912e-10	-7.255733	1.474820e-12

From the plot, we can see that the polynomial model fits well to the data, capturing the changing trend with distinct critical points. Further, we can see from the summary that all coefficients in the model are significant, indicating that all coefficients are relevant to explaining the variation in the data. The following formula was used to fit the model:

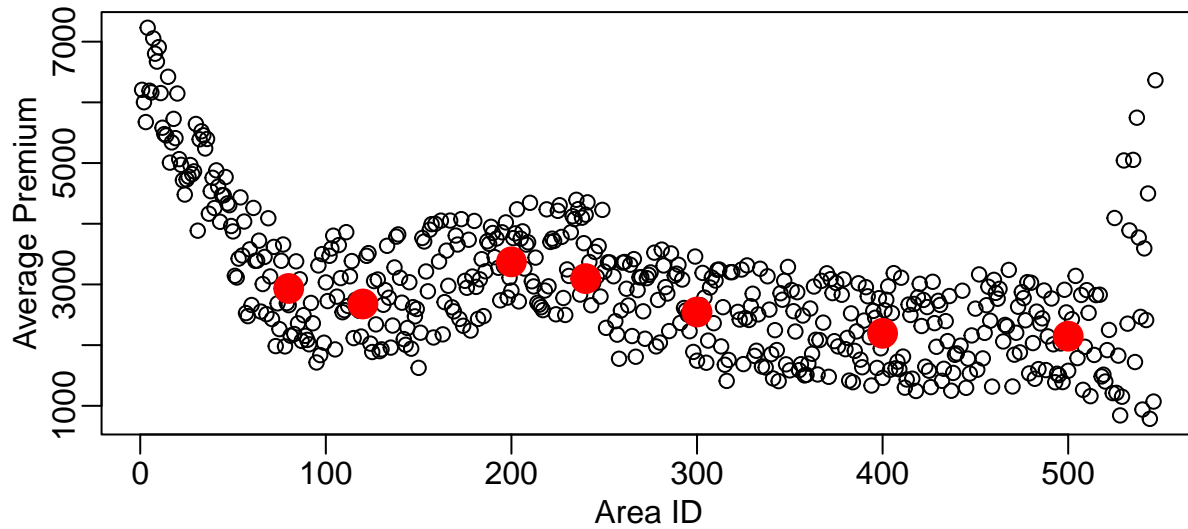
$$\hat{P} = -(3.83 \times 10^{-9})x^5 + (6.25 \times 10^{-6})x^4 - (3.71 \times 10^{-3})x^3 + (9.71 \times 10^{-1})x^2 - (1.09 \times 10^2)x + (7.18 \times 10^3)$$

Where  $\hat{P}$  is the graduated average premium, and  $x$  represents a particular Area ID.

#### 3. b. Natural Cubic Spline

By visual inspection of the training dataset scatter plot, knots were chosen at `Area.ID = (80, 120, 200, 240, 300, 400, 500)`. A plot is shown below with the chosen knots in large red circles.

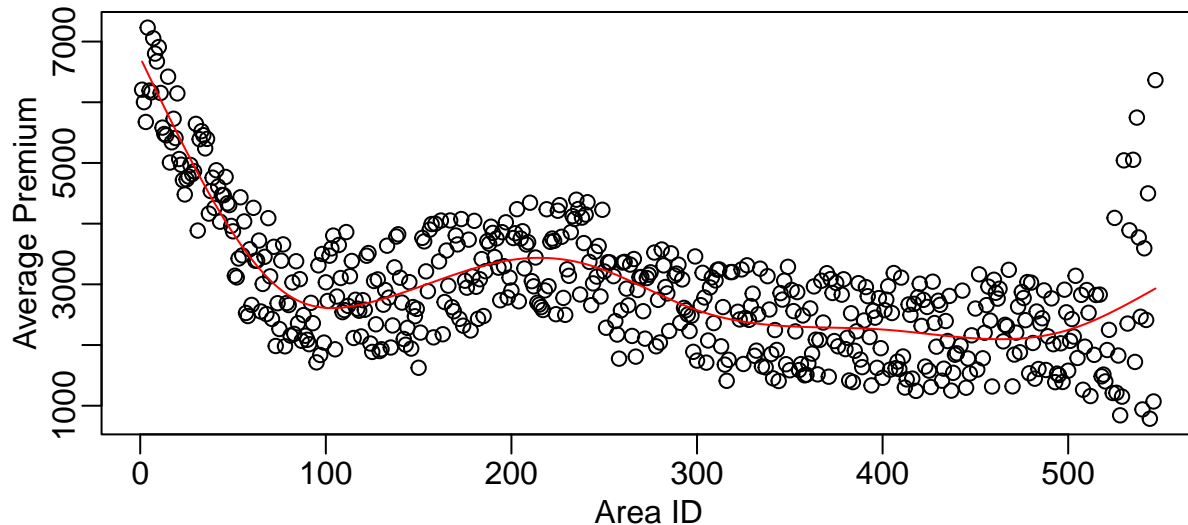
## Knots for Natural Cubic Spline



The knots have been placed where significant changes in the trend occur. Knots with this placement will help to form the base of a model that is flexible enough to capture the changes in trend but not so flexible that it is overfit to the data.

To fit the natural cubic spline to the data with the chosen knots, a loop was performed, fitting  $(2^7 - 1)$  natural cubic splines with every different possible combination of knots (excluding the combination containing no knots). Each model was compared to the validation set, using mean squared error as the performance measurement. The model with the lowest mean squared error was chosen to be the optimal model. From the loop, it was found that the best performing model was the one with knots at (80, 120, 200, 240, 300, 400, 500). This model includes all of the original knots. The model was fitted against the training dataset, shown below.

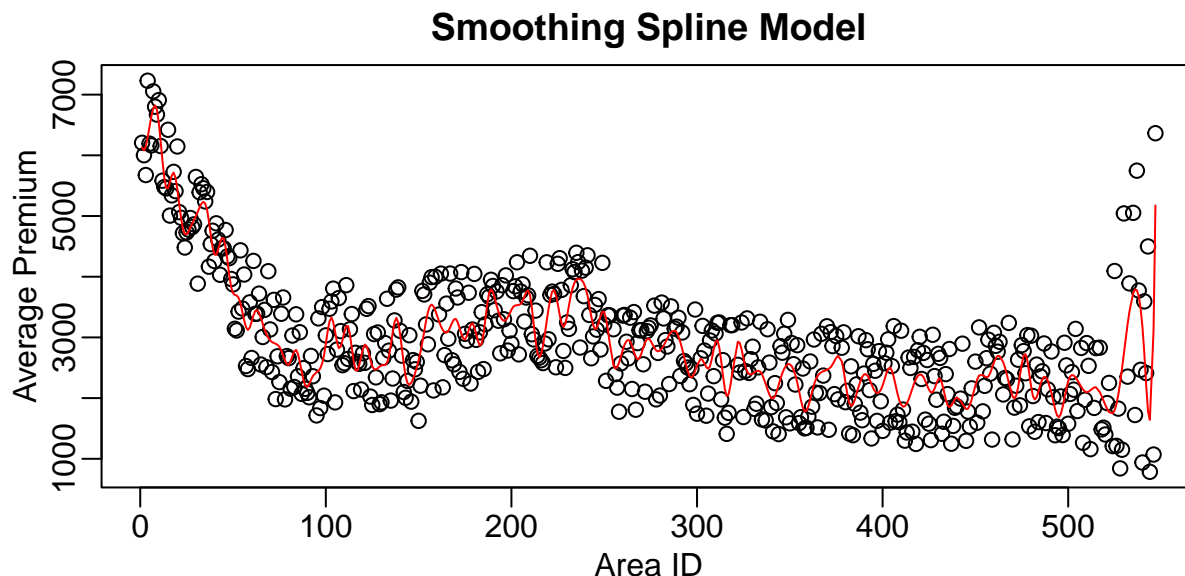
## Natural Cubic Spline Model



### 3. c. Smoothing Spline

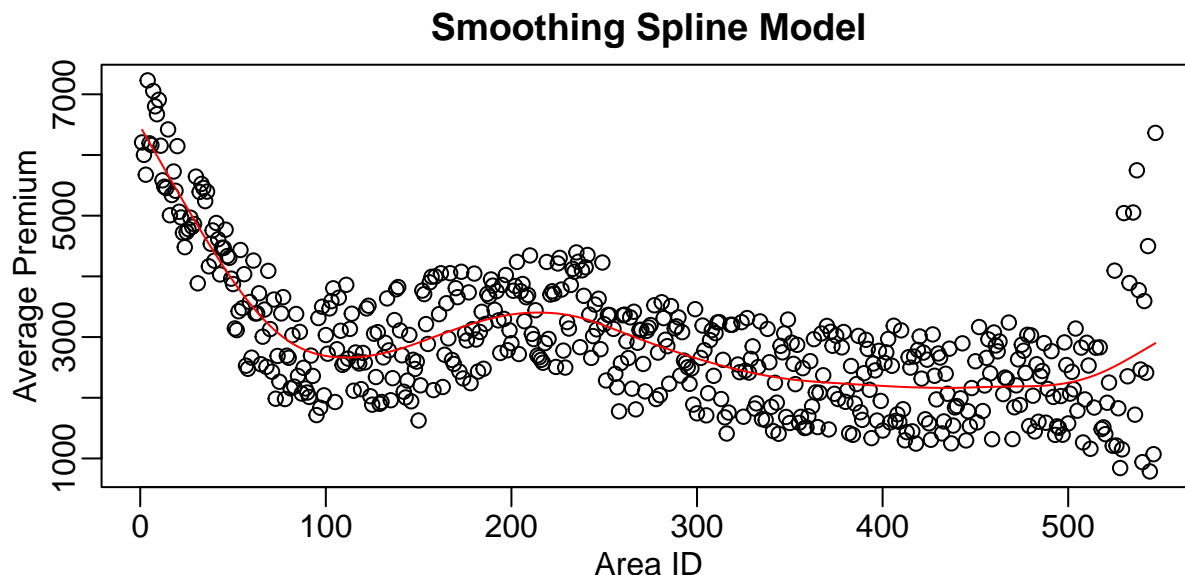
A smoothing spline was fit to the training data, using a loop to find the optimal value for the trade-off parameter ( $\lambda$ ). Within the loop, different values of  $\lambda$ , ranging from 0 to 0.01, were used to fit the model against the training set. A prediction set was made based off the model, and the fit of the model was assessed against

the validation set using mean squared error. The values of each tested lambda and the respective mean squared error from each model was stored in a data frame. From this, the optimal lambda was chosen as the lambda that produced the model with the lowest mean squared error. The optimal lambda was found to be  $\lambda = 0$ . Using this lambda to fit a new model against the training set yielded the following model:



From visual inspection, this spline is clearly overfit, as it jumps sporadically from one side of the trend to another, trying to capture every single data point. This may be a result of the validation model being too similar to the training data, especially in the case of the location of the clusters, as this will significantly impact the mean squared error calculation, resulting in an overfit model being chosen as the optimal model. After conducting further analysis, it was found that  $\lambda = 0$  produced a global minimum mean squared error, suggesting it was the true optimal model. To account for the overfitting, we will have to use a different method to find a model.

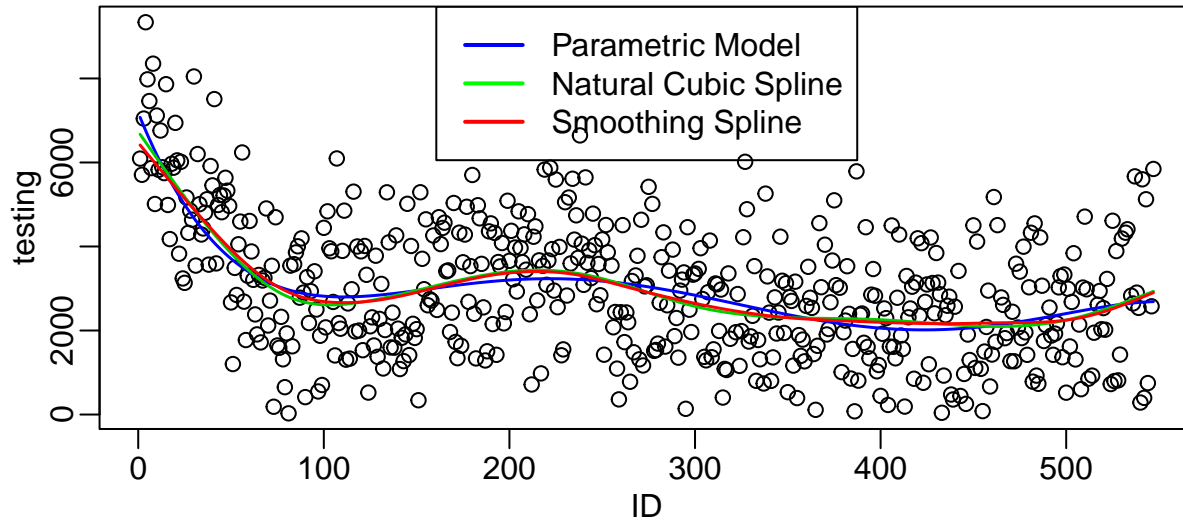
Now, we will fit a model using the validation set instead of the training set, obtaining the optimal lambda from this model by setting `cv = TRUE` within the `smooth.spline()` function. This technique uses cross-validation to determine the optimal lambda for the model. Using this value for lambda, we will then fit a new smoothing spline model to the training set. We will use this as our chosen smoothing spline model. The new model produces the following plot:



Our new lambda value is 0.0014011. Visually, this model is clearly not overfitting to the data, when compared to the model with  $\lambda = 0$ . We will use this as our optimal smoothing spline model.

### 3. d. Comparison

First, we will make a visual comparison of each model against the testing data:



All three models appear to be very similar to each other, with some slight deviations at particular points. This shows all models effectively capture the changing trends in the data.

It is often the case that the optimal smoothing spline is a cubic spline. Natural cubic splines also involve fitting cubic splines between points. This makes the two spline models extremely similar. Smoothing splines tend to be more flexible than natural cubic splines due to the adjustability of the smoothing parameter, where the natural cubic spline does not have such parameter. However, due to their complexity, smoothing splines are often more difficult to interpret than natural cubic splines.

The fitted parametric model is of degree 5, which is higher than the cubics within the spline models, which are of degree 3. This model would produce a very similar result to fitting a single spline of degree 5 between two knots at the extreme ends of the data. Due to this, parametric models tend to be less flexible than natural cubic splines, which reduces chances of overfitting, but often comes at the cost of performance.

All three models appear to accurately capture the changing trends in the data, without any absurd overfitting. We will now compare each model quantitatively, by computing the mean squared error of each model relative to the testing data.

##	Model	MSE
## 1	Parametric	1757543
## 2	Natural Cubic Spline	1715907
## 3	Smoothing Spline	1725940

The natural cubic spline is the model that produces the lowest mean squared error, making it the best model to use for this application. We will now conduct graduation tests to determine the goodness of fit of this model.

## 4. Graduation Tests

For all tests, a significance level of 0.05 (5%) has been chosen. All tests have Hypotheses:

- $H_0$ : The observed data follows the model
- $H_1$ : The observed data does not follow the model

### 4. a. i. Chi Squared Test

To perform the Chi Squared test on the graduation, we first find the expected values of the average premium, using our model's fitted values. The observed values are the training set. Since we do not know the underlying distribution of the residuals, we will standardise them based on the empirical standard deviation of the observed residuals, making the assumption that they follow a standard normal distribution from there:

$$z_x = \frac{Observed - Expected}{\sqrt{Var(Observed - Expected)}} \sim N(0, 1)$$

The Chi Squared test statistic is as follows:

$$\tau = \sum_{i=1}^n z_i^2 \sim \chi_{n-r-1}^2$$

In this case, since we chose the natural cubic spline as our best model, the degrees of freedom of the chi-squared test statistic is the number of observed values, minus the number of knots used in the model (in this case, 7), minus 1. Hence, the degrees of freedom in the chi squared test is 514.

Computing the observed test statistic of 521, we attain a p-value of 0.4058971. This value is greater than our chosen significance level of 0.05. Hence, we retain the null hypothesis, indicating that by the chi squared test, there is insufficient evidence to conclude the model is not a good fit to the data. We conclude that the model is a good fit.

### 4. a. ii. Standardised Deviations Test

The standardised deviations test involves computing the standardised deviation,  $z_x$ , (shown above) for each ID value and assigning them into buckets based on their value. For this test, the following buckets were chosen:

$$(-\infty, -3), \quad (-3, -2), \quad (-2, -1), \quad (-1, 0), \quad (0, 1), \quad (1, 2), \quad (2, 3), \quad (3, \infty)$$

The results of the analysis described above is displayed in the following table:

##	int.low	int.high	int.freq	int.expected
## 1	-Inf	-3	1	0.7046468
## 2	-3	-2	5	11.1709221
## 3	-2	-1	85	70.9424737
## 4	-1	0	172	178.1819574
## 5	0	1	165	178.1819574
## 6	1	2	88	70.9424737
## 7	2	3	2	11.1709221
## 8	3	Inf	4	0.7046468



`int.low` and `int.high` represent the lower and upper bounds of each bucket, respectively. `int.freq` is the observed number of  $z_x$ 's in each bucket. `int.expected` is the expected number of  $z_x$ 's in each bucket, based on the assumption that each  $z_x$  follows a standard normal distribution.

The test statistic for the hypothesis test follows a chi squared distribution with degrees of freedom equal to the number of buckets, minus 1. It is computed as:

$$\tau = \sum_{i=1}^n \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} \sim \chi_{n-1}^2$$

Note that we are assessing the observed and expected amounts in each bucket, rather than the data itself. After performing the test, the observed test statistic was found to be 34.5493057, resulting in a p-value of 0.0000136. Since this falls below our chosen significance level, we reject the null hypothesis. By the standardised deviations test, there is sufficient evidence to suggest that the chosen model does not fit the data.

#### 4. a. iii. Signs Test

The signs test involves observing the number of points that are above and below the graduated line. If the graduation is a good fit, there should be a roughly equal amount of points above and below the line. Let  $P$  be the number of points above the line. If the model is a good fit,  $P$  should be distributed as:

$$P \sim \text{Binomial}(m, 0.5)$$

Where  $m$  is the total number of observations. Since we are dealing with a large sample size, we can use the central limit theorem to assume a normal distribution of the number of positive values:

$$P \sim N(0.5m, 0.25m)$$

Conducting a two-sided statistical test based off this assumption yields a test statistic of -0.1750752, and a p-value of 0.8610205. Since our observed p-value is far greater than the significance level of 0.05, we retain the null hypothesis, concluding that the model is a good fit by the signs test.

#### 4. a. iv. Cumulative Deviations Test

The cumulative deviations test involves assessing the fit if a model relative to cumulative residuals. It aims to address the issue of detecting large deviations that the chi-squared test is unable to do. The test statistic is shown below:

$$\tau = \frac{\sum_{i=1}^n (\text{Observed}_i - \text{Expected}_i)}{\sqrt{\sum_{i=1}^n (\text{Expected}_i)}} \sim N(0, 1)$$

Conducting this test on our model, we found test statistic of 0, and a p-value of 1. The p-value is larger than the chosen significance level of 0.05. Thus, we retain the null hypothesis. There is insufficient evidence to suggest the model does not fit the data, according to the cumulative deviations test.

#### 4. a. v. Grouping of Signs Test

The grouping of signs test involves observing groups of positive deviations, in order to determine whether the data has been overgraduated. This test aims to combat the shortcomings of the signs test, as a model that is too smooth can still pass the signs test, even though it may not be a suitable model for the data.

Let  $G$  be the number of groups of positive signs. If the graduation has been done properly, we will expect  $G$  to be large. Let  $m$  be the number of deviations,  $n_1$  be the number of positive deviations and  $n_2 = m - n_1$  be the number

of negative deviations. Under the null hypothesis, we assume  $G$  follow a hypergeometric distribution, with the following PMF:

$$P(G = g) = \frac{\binom{n_1-1}{g-1} \binom{n_2+1}{g}}{\binom{m}{n_1}}$$

Since we dealing with a large data set, we can use the following normal approximation for  $G$  to perform our test:

$$G \sim N\left(\frac{n_1(n_2+1)}{n_1+n_2}, \frac{(n_1n_2)^2}{(n_1+n_2)^3}\right)$$

Conducting a one-sided statistical test (that rejects the null hypothesis if  $G$  is too small) based off this approximation yielded a test statistic of -0.173073, with a p-value of 0.431297. The observed p-value is greater than the chosen significance level of 0.05, so we retain the null hypothesis, confirming that our model has not been overgraduated by the grouping of signs test.

#### 4. a. vi. Serial Correlations Test

If we have performed our graduation properly, the standardised deviations should all be independent of each other. If our model is over-graduated, we should observe some correlation between the deviations. The two sequences

$$z_1, z_2, \dots, z_{m-1} \quad \text{and} \quad z_2, z_3, \dots, z_m$$

Should be uncorrelated to each other. Let  $r$  be the correlation between the two sequences. Under the null hypothesis, we find that

$$r\sqrt{m} \sim N(0, 1)$$

The correlation between the two sequences was found to be  $r = -0.0229406$ , resulting in a p-value of 0.6999063. The p-value is greater than the significance level of 0.05. Therefore, there is insufficient evidence to reject the null hypothesis. Thus, we retain the null hypothesis, and by the serial correlations test, the model is not over-graduated.

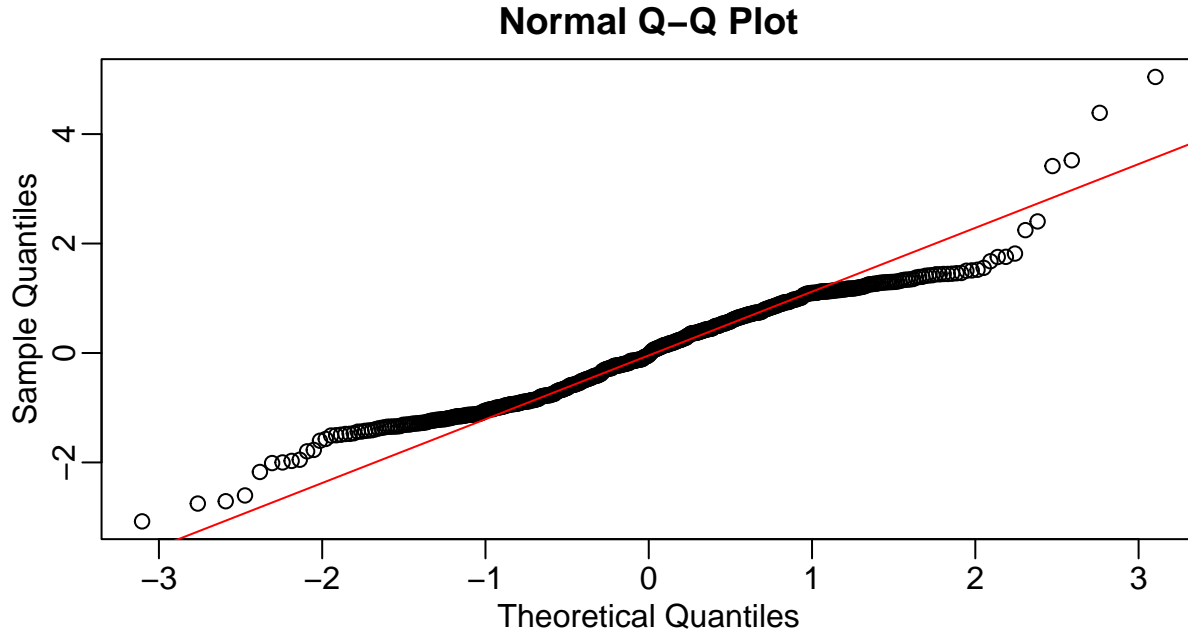
#### 4. a. vii. Graduation Test Summary

A table summarising the results from our test are below:

##	Test	p.Value	Retain_H0
## 1	Chi-Squared Test	0.40589714308	TRUE
## 2	Standardised Deviations Test	0.00001359048	FALSE
## 3	Signs Test	0.86102051039	TRUE
## 4	Cumulative Deviations Test	1.00000000000	TRUE
## 5	Grouping of Signs Test	0.43129700515	TRUE
## 6	Serial Correlations Test	0.69990630439	TRUE

The grouping of signs and serial correlations tests both test for over-graduation. Our model passed both these tests, meaning it is clear that our model has not been over-graduated (i.e., not too smooth).

On the other hand, our model failed the standardised deviations test. We'll have to investigate this further. A QQ-plot of the standardised deviations is shown below:



The QQ-plot appears to have significant deviations from the normal line, suggesting that the underlying standard normal distribution assumption of the standardised deviations may not be valid. A solution to this would be to conduct further analysis to determine the underlying distribution of the standardised deviations.

The underlying assumptions tied to the standardised deviations rely on the variance to remain constant throughout the entire set. Clearly, from the scatterplots shown in the Preliminary Data Analysis section, the data appears to go through phases of varying variance, exhibiting a property known as heteroskedasticity. This property may be the reason for the models' under-performance in the standardised deviations test. Finding a method of conducting this test without requiring the underlying assumption of constant variance of the standardised deviations may solve this issue.

#### 4. b. Model Shortcomings

Natural cubic splines have some shortcomings. One of the biggest issues is that they require a subjective choice of knots. If sub-optimal knots are chosen, it may lead to an inaccurate, poor fit to the data. Too few knots can lead to the spline under-fitting, as it will be unable to capture the true nature of trends within the data. Too many knots can lead to the spline over-fitting, as it will try to capture random noise that does not represent the underlying trend.

Further, natural cubic splines assume a linear relationship before the first knot and after the last knot. In some applications, this may not be appropriate, as there could be some non-linear trend within for these points that the spline will not capture. A solution to this would be to place knots at the extreme ends of the data. In the case of the average premiums dataset, there appears to be a distinct linear trend for Area ID's less than 80, so this should not be a problem for this set.

The complexity of a natural cubic spline makes it more difficult to interpret than other, simpler models, especially with splines with a larger number of knots. Poor interpretability can lead to difficulties in communicating results, especially to those who don't have an actuarial or statistical background. In the context of home insurance affordability, using natural cubic splines to explain trends found in the data may make it difficult to communicate insights, as non-specialists may struggle to understand the implications of the model in mapping and predicting policy costs.

## 5. Model Discussion

### Implications of the Model

The natural cubic spline model is able to capture the general trend of the data well.

This model could be used to capture the effect of adding bushfire risk to the existing CRP, noting any differences from the original model (used on the data in this assignment) to a new model (using data gathered some time after imposing the new CRP approach). If the new model predicts significantly lower average premiums than the old model at every single Area ID, then we can say adding bushfire risk to the CRP had a significant effect in reducing premiums. We should also be looking for a reduction in the variation of premiums, since the new reinsurance pool will better diversify risk, allowing for premiums to be better smoothed out among all Area ID's. This should be reflected by fewer sharp changes in the direction/concavity in the data and in the new model.

However, some time after the changes to the CRP have been made, if the new model has shown only a marginal decrease in premiums, little change in the variation, and little change to the shape of the data, it may suggest that adding bushfire risk to the existing CRP had little impact on reducing premiums, and may not be a sufficient solution to improving the home insurance affordability issue. We could conclude that the new approach failed to do as it was intended, and a new solution to the home insurance affordability issue will need to be considered.

### Potential Improvements to the Model

The natural cubic spline has some potential for improvement.

Finding better knot placements through the use of a mathematically rigorous method can help to choose knots that are able to reflect trends in the data better. An example of this is to select knots at all integer x-values, and one-by-one, remove knots that add little to no information until an optimal model is found. Another method could be to use a penalised regression technique such as LASSO optimisation, which removes knots by applying a penalty to the coefficients of the spline function, excluding less important knots from the model.

Exploring other methods for evaluating performance (e.g., R-squared, AIC, BIC, weighted MSE, etc.) could be used to find a better model. Using only mean squared error to gauge performance of a model can be quite limiting, as it only considers the average squared residuals, which may not show the full picture of the data. It is possible that using mean squared error can reward models that are overfit to the data, resulting in a sub-optimal model being chosen as the optimal model. Mean squared error can fail to capture specific trends within the data, and depending on the context, may lead to inappropriate models being used for the application. Using a combination of methods for performance evaluation can aid in the creation of a more robust, well-fit model that is capable of capturing underlying trends with greater accuracy, allowing the model to better align with the specific needs of the application.

For natural cubic spline graduation, as explained earlier, improvements can be made to the linearity assumption for points before the first knot and beyond the last knot. Adding knots at the extreme ends of the data will help in capturing non-linear trends outside of the range of knots used. This can then be used to accurately model trends within these areas of the data.

For parametric graduation, exploring functions outside that of just polynomials could be useful. Adding exponential, logarithmic, trigonometric, or other functions to the parametric graduation can enable the model to capture trends that our polynomial model was unable to capture. However, this can easily lead to the model being overfit to the data, making it perform poorly in testing, resulting in inaccuracy in modelling and predicting future data.

For smoothing spline graduation, optimising for a non-zero trade-off parameter will enable the model to perform better. Imposing restrictions on the values that the trade-off parameter can take upon testing performance, as well as using a combination of other performance metrics to determine performance of the model may help in finding a better model that is not overfit to the data.