In this work we simulated a function \( q(x) \) given by:
\[
\frac{\sin(x)}{x} + 0.5 \cos(2x + \frac{\pi}{4}).
\] (1)

We then added some random normally distributed noise with a mean of zero and a standard deviation of 0.1 to our grid of values. Next, we used a variety of methods to obtain an estimate \( \hat{f}(x) \) for the true function \( q(x) \) based on simulated data. These methods required the use of the mgcv package and the freeknotspline package. We made use of bootstrapping in conjunction with both of these.

### What is a Spline?

Regressions splines are used to construct a model \( \hat{f}(x) \) to fit a set of data. Many of them require knots to do this.

Regression splines are a linear combination of basis function which depend on a set of knot points. The basis functions are constructed such that the resulting linear combination will be continuous and have a certain number of continuous derivatives. The basis functions highly depend on the placement of knot points.

In this work we mainly used penalised splines (or P-splines) which punish models which vary too rapidly. To fit this model we sought to minimise the following equation:
\[
||y - X\beta||^2 + \lambda \int |f''(x)|^2 \, dx
\] (2)

where \( X \) is our model matrix, \( \beta \) is a vector of unknown parameters that we are trying to find and \( \lambda \) is a tuning parameter [1].

### Analysis & Results

We ran two different simulations in the project.

Firstly we used bootstrapping with the mgcv package to find our fits and confidence intervals. The fit obtained is shown in Figure 1.

In the second method we ran the genetic algorithm to find the best set of knots for our data set. We then used these knots to fit to all the bootstrapped data sets. The fit obtained from this is shown in Figure 2.

We assessed the quality of the fits by considering the mean squared error given by:
\[
\frac{1}{n} \sum_{i=1}^{n} (\hat{f}(x_i) - q(x_i))^2.
\] (3)

where the \( x_i \) are our grid points. The results are shown in Table 1.

In the future we would like to try running the genetic algorithm for each bootstrapped data set to find their individual set of optimum knots.

### References


### Optimizing Knot Placement in Regression Splines

Joseph Holey  Michael O’Malley

STOR-i, Lancaster University

---

**Table 1**: This table shows the Mean Squared Error for the two methods between the truth and the fit.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Knots</td>
<td>(5.896 \times 10^{-4})</td>
</tr>
<tr>
<td>Genetic algorithm</td>
<td>(5.762 \times 10^{-4})</td>
</tr>
</tbody>
</table>