Analysing Engine Wear and Size with Cubic Smoothing Splines

Student Name: Yiwei Yang

StudentID: 201749736

Submitted Time: 2024-04-16

1 Introduction

This report investigates the relationship between engine wear and size by doing graphical exploration and fitting smoothing spline models. Our focus lies in assessing the impact of the smoothing parameter λ on model fitting and parameter selection. By analyzing how different λ values impact model fitting and preserving monotonicity, we aim to provide insights into effective parameter selection strategies for optimal model performance.

2 Data Exploration

Firstly, we visualised the data and explored the potential relationships between variables. Figure 1 shows a strong correlation(approx. -0.62 from `cor.test` function in `R`) between car engine size and wear. That is, a bigger engine size is associated with a higher wear index. This graph might indicate that there is a negative correlation between wear and size. However, the data suggest that a cubic spline model is more appropriate, as evidenced by the non-linear fit line.

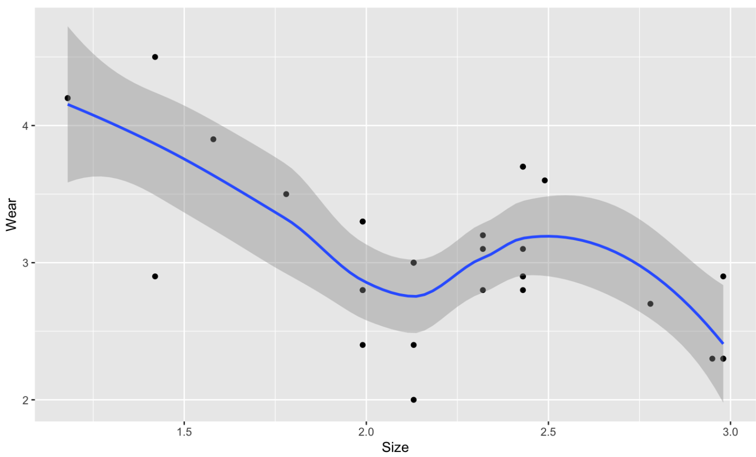

Figure 1: The Relationship Between Engine Wear and Size.

3 Fitting Cubic Smoothing Splines

In this section, we will investigate the relationship between engine car wear and size by fitting cubic smoothing spline models with different values of the smoothing parameter(λ) and discuss the effect of λ on the smoothness of the fitted curves. Figure 2 shows that as the smoothing parameter λ increases, the model fitting line

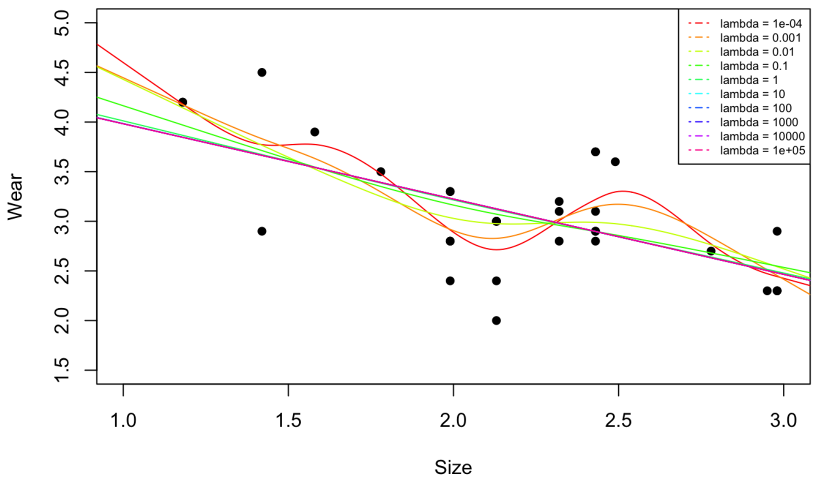

Figure 2: Smoothing Spline Model Fitting with different parameters(λ).

progressively becomes a monotonic straight line from a cubic spline. In addition, as λ approaches 0, the model tends to fit as much as the sample data, leading to overfitting and limiting its predictive capability on unseen data. Conversely, as λ approaches infinity, the model simplifies to a least squared fit, resulting in poor data interpretation due to large residuals.

In the two extreme cases, the function of the optimal solution ranges from very smooth to rough. We hope that we can obtain an appropriate model by changing the value of $\lambda \in (0, \infty)$.

4 Evaluation of Model Fit

We introduce the Generalised Cross-Validation(GCV) to select a suitable range for λ and choose an optimal value of λ. With this approach, a lower GCV criterion means a better fitting.

Figure 3 shows the Generalised Cross-Validation(GCV) approach criterion with different parameter λ. We see that the GCV criterion is decreasing with a small value of λ, reaching its lowest point at 0.21 with λ of 0.0015. Then, the GCV criterion gradually increases and stabilizes of 2.25 when λ reaches 0.012, meaning the model fitting tends to be a straight line.

In conclusion, we can choose 0.0015 as the best value of the smoothing parameter λ. In addition, any value of $\lambda \in (1e-04, 1e-02)$ is preferable for model fitting since the responding criterion of GCV is relatively low.

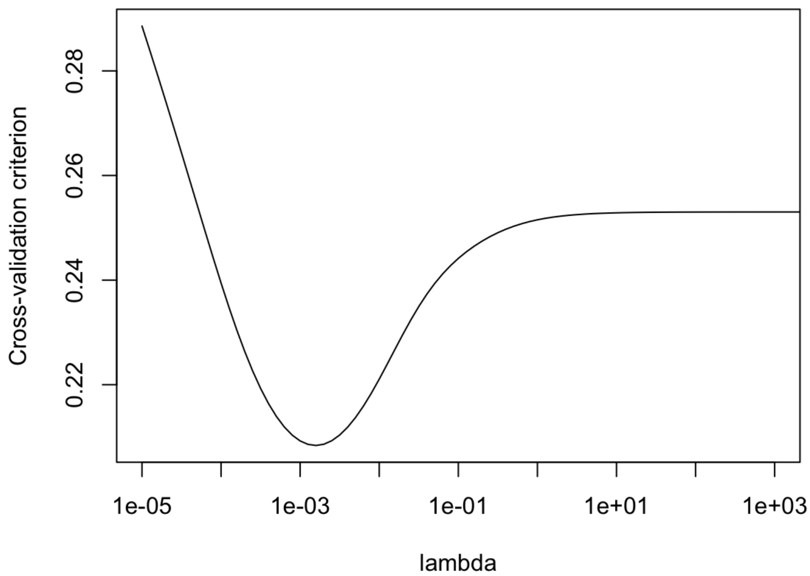

Figure 3: The Generalised Cross-Validation(GCV) values with different smoothing parameters λ.

5 Prediction and Sensitivity Analysis

In this section, we introduce the Sensitivity Analysis by fitting the model with different parameters λ to predict the wear of a car with engine size of $2.6L$. This method helps us to evaluate the model performance under varying parameter values and identify the optimal parameter value.

Figure 4 shows the change of predicted wear values as λ varies. The predicted value decreases from 3.26 to 2.77 and stabilizes at 2.77 as λ approaches approximately

1.70. This stabilization of the predicted value suggests that the model tends towards linear regression, resulting in a gradual reduction in the data fit and less significant changes as predicted values. On the contrary, when we choose $\lambda \in (1e-04, 1e-02)$, the model is sensitive to changes, which can be interpreted as the λ we chose has a significant effect on model fitting. In conclusion, models become less effective for λ values exceeding 1.70.

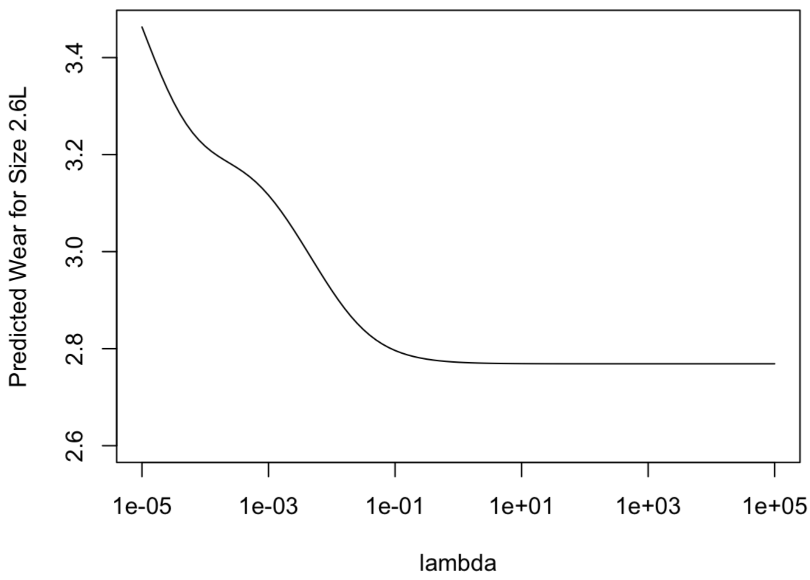

Figure 4: Predicted Values of Size 2.6L with different parameters λ

6 Conclusions

In summary, our analysis highlights the importance of parameter selection when fitting smoothing spline models to investigate the relationship between car engine wear and size. Through graphical exploration and model fitting, we observed the model fitting become very smooth from rough as λ increases. The Generalised Cross-Validation approach identified the optimal value of $\lambda = 0.0015$ for model fitting, while sensitivity analysis double-checked that the value of λ we choose is appropriate.

7 Appendix

```r
# Spline Line
splinedata<- read.csv(
  "https://rgaykroyd.github.io/MATH3823/
  Datasets/engine-36.csv", header=T)
head(splinedata)
cor(splinedata)
# exploring the data
x <- splinedata$size
y <- splinedata$wear
ggplot(splinedata, aes(x = x,y = y)) +
  geom_point() +
  geom_smooth()+
  labs(x = "Size", y = "Wear")
summary(splinedata)

# modelling
myfit1 = smooth.spline(
  splinedata$size,splinedata$wear,lambda = 0.01)
curve(myfit1(x,deriv=1), 0, 10, lwd=1.5,
      xlab="x", ylab="First derivative")
abline(v=x, col="grey"); abline(h=0, col="grey")

residuals <- splinedata$wear - predict(
  myfit1, x = splinedata$size)$y
# Calculate standard deviation of residuals
residuals_sd <- sd(residuals)
# Output standard deviation of residuals
print(residuals_sd)

# visualise mtfit
plot(splinedata$size,splinedata$wear,
     xlim=c(1,3),ylim=c(1,5),pch=16)
fit.locations = seq(0,10,0.01)
fitted = predict(myfit1, fit.locations)
lines(fitted,col="blue")


# Create a vector of lambda values
lambda_values <- 10^seq(-4,5,by=1)
# Create an empty plot
```

```r
plot(splinedata$size, splinedata$wear,
     xlim = c(1, 3), ylim = c(1.5, 5),
     pch = 16, xlab = "Size", ylab = "Wear")
results <- list()
for (lambda_val in lambda_values) {
  myfit <- smooth.spline(splinedata$size,
                         splinedata$wear,
                         lambda = lambda_val)
  fitted <- predict(myfit, fit.locations)
  lines(fitted,
        col = rainbow(length(lambda_values))[
          which(lambda_values == lambda_val)])
  lam <- lambda_val
  cv <- myfit$cv.crit

}
# Add legend
legend("topright",
       legend = paste("lambda =", lambda_values),
       col = rainbow(
         length(lambda_values)), lty = 4, cex = 0.6)
print(cv)

# Create a vector of lambda values
lambda_values <- 10^seq(-5, 5, by = 0.1)
plot(splinedata$size, splinedata$wear,
     xlim = c(1, 3), ylim = c(1.5, 5),
     pch = 16, xlab = "Size", ylab = "Wear")
# Fit smoothing splines with different lambda values
results <- list()
cv_values <- numeric(length(lambda_values))
for (i in seq_along(lambda_values)) {
  lambda_val <- lambda_values[i]
  myfit <- smooth.spline(
    splinedata$size, splinedata$wear, lambda = lambda_val)
  fitted <- predict(myfit, fit.locations)
  # Plot the fitted curve
  lines(fitted, col = rainbow(length(lambda_values))[i])
  print(lambda_val)
  print(myfit$cv.crit)
  # Store cross-validation value
  cv_values[i] <- myfit$cv.crit
```

```
}
legend("topright",
       legend = paste("lambda =", lambda_values),
       col = rainbow(length(lambda_values)),
       lty = 1, cex = 0.6)
# Plot cross-validation values
plot(lambda_values, cv_values,
     type = "l", xlab = "lambda",
     ylab = "Cross-validation criterion",
     xlim = c(1e-05, 1e+03),log = "x")

# predicted values
predicted_values <- numeric(length(lambda_values))

size_to_predict <- 2.6
for (i in seq_along(lambda_values)) {
  lambda_val <- lambda_values[i]
  myfit <- smooth.spline(
    splinedata$size, splinedata$wear, lambda = lambda_val)
  predicted_values[i] <- predict(myfit, size_to_predict)$y
  print(lambda_val)
  print(predict(myfit, size_to_predict)$y)
}

plot(lambda_values, predicted_values,
     type = "l", xlab = "lambda",
     ylab = "Predicted Wear for Size 2.6L",
     ylim = c(2.6, max(predicted_values)), log = "x")
```

UNIVERSITY OF LEEDS

School of Mathematics

Declaration of Academic Integrity
for Individual Pieces of Work

I declare that I am aware that as a member of the University community at the University of Leeds I have committed to working with Academic Integrity and that this means that my work must be a true expression of my own understanding and ideas, giving credit to others where their work contributes to mine.

I declare that the attached submission is my own work.

Where the work of others has contributed to my work, I have given full acknowledgement using the appropriate referencing conventions for my programme of study.

I confirm that the attached submission has not been submitted for marks or credits in a different module or for a different qualification or completed prior to entry to the University.

I have read and understood the University's rules on Academic Misconduct. I know that if I commit an academic misconduct offence there can be serious disciplinary consequences.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties to verify that this is my own work, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and I wish to have taken into account.

Student Signature: Yiwei Yang **Student Number:** 201749736

Student Name: Yiwei Yang **Date:** 16/04/2024

Please note:

When you become a registered student of the University at first and any subsequent registration you sign the following authorisation and declaration:

"I confirm that the information I have given on this form is correct. I agree to observe the provisions of the University's Charter, Statutes, Ordinances, Regulations and Codes of Practice for the time being in force. I know that it is my responsibility to be aware of their contents and that I can read them on the University web site. I acknowledge my obligation under the Payment of Fees Section in the Handbook to pay all charges to the University on demand.

I agree to the University processing my personal data (including sensitive data) in accordance with its Code of Practice on Data Protection http://www.leeds.ac.uk/dpa . I consent to the University making available to third parties (who may be based outside the European Economic Area) any of my work in any form for standards and monitoring purposes including verifying the absence of plagiarised material. I agree that third parties may retain copies of my work for these purposes on the understanding that the third party will not disclose my identity."'