BIOS 7720: Applied Functional Data Analysis

Lecture 8: Scalar on Function Regression (SoFR)

Andrew Leroux

April 6, 2021

Roadmap

1. Logistics
2. In Class Exercises

Logistics

- Homework 1
  - Grading
  - Solutions
- Homeworks 2/3
  - Due dates:
    - HW 2: 4/26
    - HW 3: 5/6
  - Lengths
    - 4 (generally shorter) questions each
    - No “extensive" simulation studies

In Class Exercises

- Break into 4 groups
- Each group gets a different question (next slide)
- 50 minutes to work together to answer the question
- 20 minutes to present on group solutions
- Use the NHANES physical activity data created by the *R* script "lecture_8.R"

In Class Exercises: Problem 1

- Consider the numeric approximation approach to fitting GFLM

$$
\begin{aligned}
g(E[y_i|x_i, \boldsymbol{z}_i]) &= \alpha_0 + x_i\beta + \int_{\mathcal{S}} z_i(s)\gamma(s)ds \\
&= \alpha_0 + x_i\beta + \int_{\mathcal{S}} z_i(s)\sum_k \xi_k\phi_k(s)ds \\
&\approx \alpha_0 + x_i\beta + \sum_{j=1}^{J} l(s_j)z_i(s_j)\left[\sum_{k=1}^{K_\gamma}\xi_k^\gamma\phi_k^\gamma(s_j)\right] \\
&= \alpha_0 + x_i\beta + \sum_{k=1}^{K_\gamma}\xi_k^\gamma\left[\sum_{j=1}^{J} l(s_j)z_i(s_j)\phi_k^\gamma(s_j)\right]
\end{aligned}
$$

- Where $l(s_j)$ is the quadrature weight associated with the numeric approximation method

In Class Exercises: Problem 1

- Consider the GFLM from last class fit to the NHANES data (5-year mortality on smoothed PA profiles)
- Do the following:
  1. Using the formula from the previous slide, fit the **unpenalized** model as a standard GLM using a small number of basis functions (e.g. $K_\gamma = 5$) using Riemann integration. You can use *mgcv* to set up the relevant basis, but **don't** use the "by" argument. **Hint**: you will not want to use the matrix *smat*, you'll need to use another object which is related to *smat*.
  2. Compare your results from the previous step to those obtained from the unpenalized fit using *mgcv* by plotting point estimates with 95% pointwise CIs for $\hat{\gamma}(s)$. The *mgcv* unpenalized fit can be estimated using the syntax:

```
fit_fglm_ps <- gam(mort_5yr ~ s(smat, by=zlmat, bs="cc",k=5, fx=TRUE), c
                   method="REML", family=binomial)
```

In Class Exercises: Problem 2

- Now, we'll incorporate the penalty term
- You can use *mgcv* to perform custom regularized regression when you have a quadratic penalty
- This is performed using the "paraPen" argument
- Take for example the additive model

$$g(E[y_i|\text{TLAC}_i]) = \beta_0 + f(\text{TLAC}_i)$$

- Where

$$\begin{aligned}\text{TLAC}_i &= \sum_{j=1}^{J} \log(1 + AC_i(s_j)) \\ &= \sum_{j=1}^{J} z_i(s_j)\end{aligned}$$

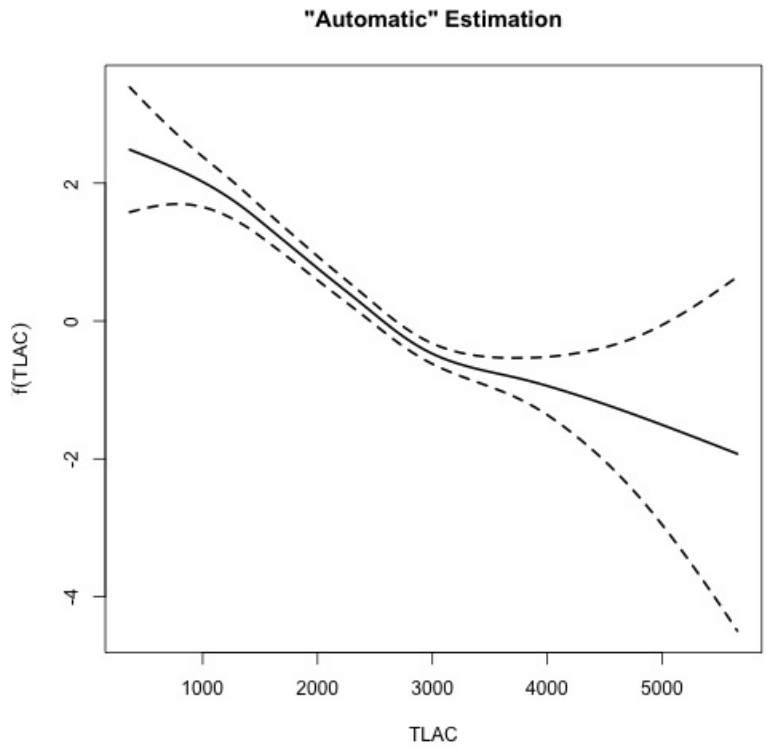
In Class Exercises: Problem 2

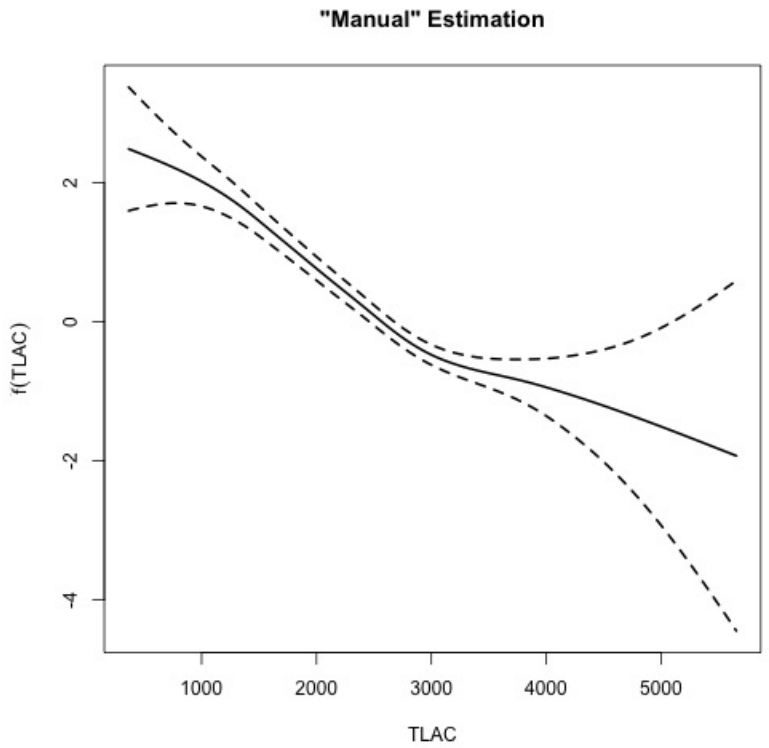
```
df$TLAC <- rowSums(df$Zraw,na.rm=TRUE) ## calculate TLAC

## let mgcv do everything
fit_TLAC <- gam(mort_5yr ~ s(TLAC, bs="cr",k=30), data=df,
                family=binomial, method="REML")

## do the fitting manually
# set up the smooth
smTLAC <- smoothCon(s(TLAC, bs="cr",k=30), data=df,absorb.cons = TRUE)
# get the basis matrix
Phi_TLAC <- smTLAC[[1]]$X
# get the penalty
S_TLAC   <- smTLAC[[1]]$S[[1]]
# fit the manual model
fit_TLAC_man <- gam(mort_5yr ~ Phi_TLAC, data=df,
                    paraPen=list(Phi_TLAC=list(S_TLAC)),
                family=binomial, method="REML")
```

In Class Exercises: Problem 2

```
## get range of TLAC values to predict over
TLAC_pred <- seq(min(df$TLAC), max(df$TLAC), len=100)
## get basis matrix for the TLAC values to predict over
Phi_pred <- PredictMat(smTLAC[[1]], data=data.frame(TLAC=TLAC_pred))
## get point estimates for \hat{f}(TLAC)
fhat_TLAC <- Phi_pred %*% coef(fit_TLAC_man)[-1]
## get the variance/covariance matrix for the spline coefficients
var_xi_TLAC <- vcov(fit_TLAC_man)[-1,-1]
## get se(\hat{f}(TLAC))
se_fhat_TLAC <- sqrt(diag(Phi_pred %*% var_xi_TLAC %*% t(Phi_pred)))
```

In Class Exercises: Problem 2





In Class Exercises: Problem 2

- Do the following:
  1. Combine your results from problem 1 with the approach presented above to fit the functional regression model

$$g(E[y_i|z_i]) = \alpha_0 + \int z_i(s)\gamma(s)ds$$

where y_i is a binary RV for 5-year all cause mortality and z_i is the smoothed log activity count profile. This model, as given last lecture, is specified by the syntax:

```
fit_fglm_ps <- gam(mort_5yr ~ s(smat, by=zlmat, bs="cc",k=30),
                   data=df, method="REML", family=binomial)
```

Again, the basis can be constructed using *mgcv::smoothCon*, but do not us the "by" argument to *s()*.

  2. Plot $\hat{\gamma}(s) + / - 2\text{SE}\hat{\gamma}(s)$ from your manual fit and from the automatic fits, show they're identical. Do not use *plot.gam()*.

In Class Exercises: Problem 3

- Do the following:
  1. Try adjusting for potential confoudning variables such as age or comorbidities included in the dataset
  2. How do the estimates of the association change?