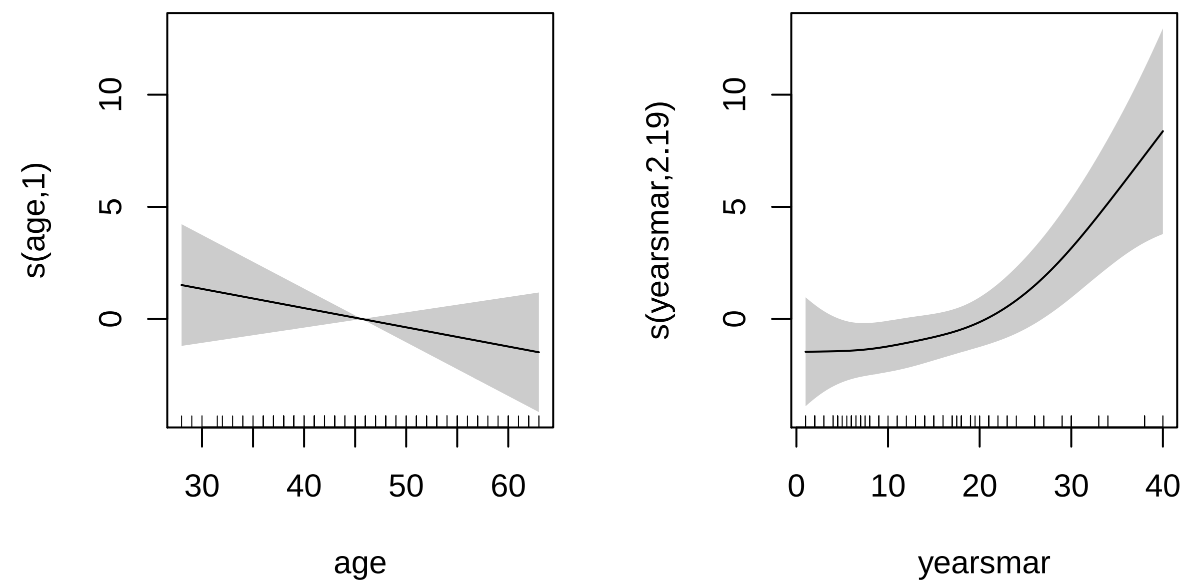
Homework 5 Solutions

Solution

29 March 2018

1 Comparing methods

- We're going to regress `happiness.scale.mean` (a measure of happiness) on `treatment` (0 is control, 1 is treatment), `female` (1 if female), `incomegt40k` (1 if income > $40,000), `degree` (1 if they have a college degree). We don't need to convert these 4 to factors since there are only two levels (though the coefficient names in `summary` would be more interpretable if we recoded them). We will also use `age` and `yearsmar`.
- Fit a linear model, call it `sslm`.
- Fit an additive model (hint: only continuous variables need to be smoothed), call it `ssgam`. Plot the partial response functions.

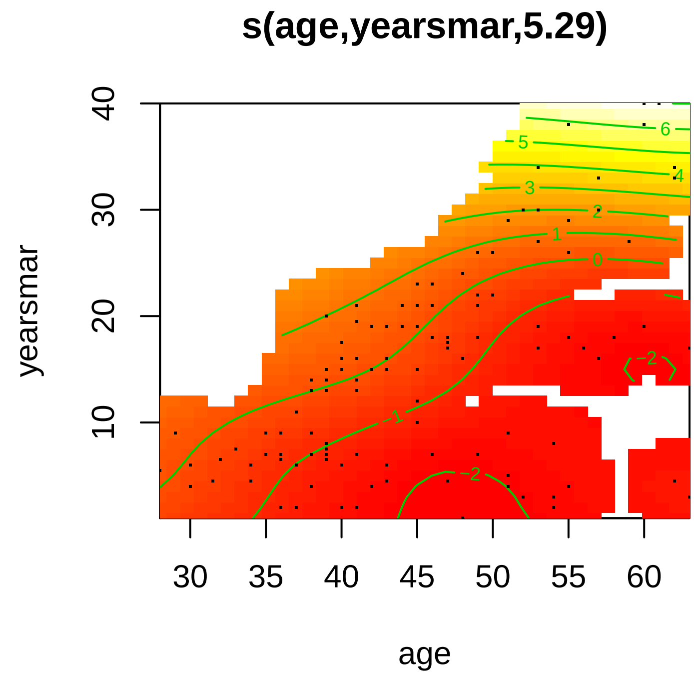


- We don't need to smooth factors because they can only take discrete outcomes. Therefore, the only possible functional dependence is additive changes from one level to the next. This is exactly what happens when you include factors linearly.
- The coefficients are -1.91 and -1.979 respectively. Both of these indicate that the treatment (doubling sexual activity) leads to a decrease in happiness (however that was measured).
- The following code extracts the leave-one-out CV from both models and prints it.

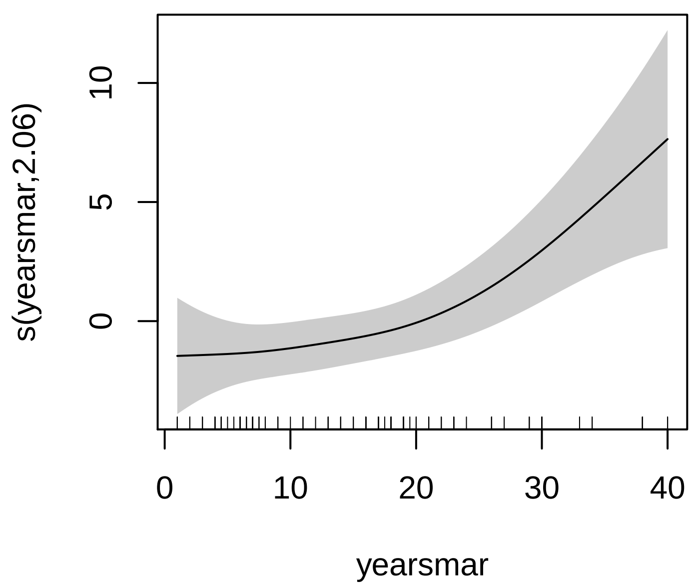
```
##   cv.glm.sslm. cv.gam.ssgam.
## 1     35.83492      34.35931
```

Based on leave-one-out CV, we slightly prefer the additive model, mainly because of the seemingly quadratic effect of `yearsmar` on happiness.

- Re-estimate the additive model but smoothing `age` and `yearsmar` together.



- Someone suggests we recode age into bins: $0-40$, $41-50$, $51-60$, $60+$. Use the `cut` function to make a new variable, `age2`. Estimate the linear model and an appropriate additive model using `age2` instead of `age`.



It is no longer useful to smooth `age2` since it is a factor (see above).

- You have now made 3 more models. Modify the code to calculate their leave-one-out CV scores. Which model is better?

```
##   cv.glm.sslm. cv.glm.sslm2. cv.gam.ssgam. cv.gam.ssgam2. cv.gam.ssgam3.
## 1     35.83492      37.06541      34.35931       34.86067       36.04815
```

Based on this output, it seems like the original additive model is preferred.

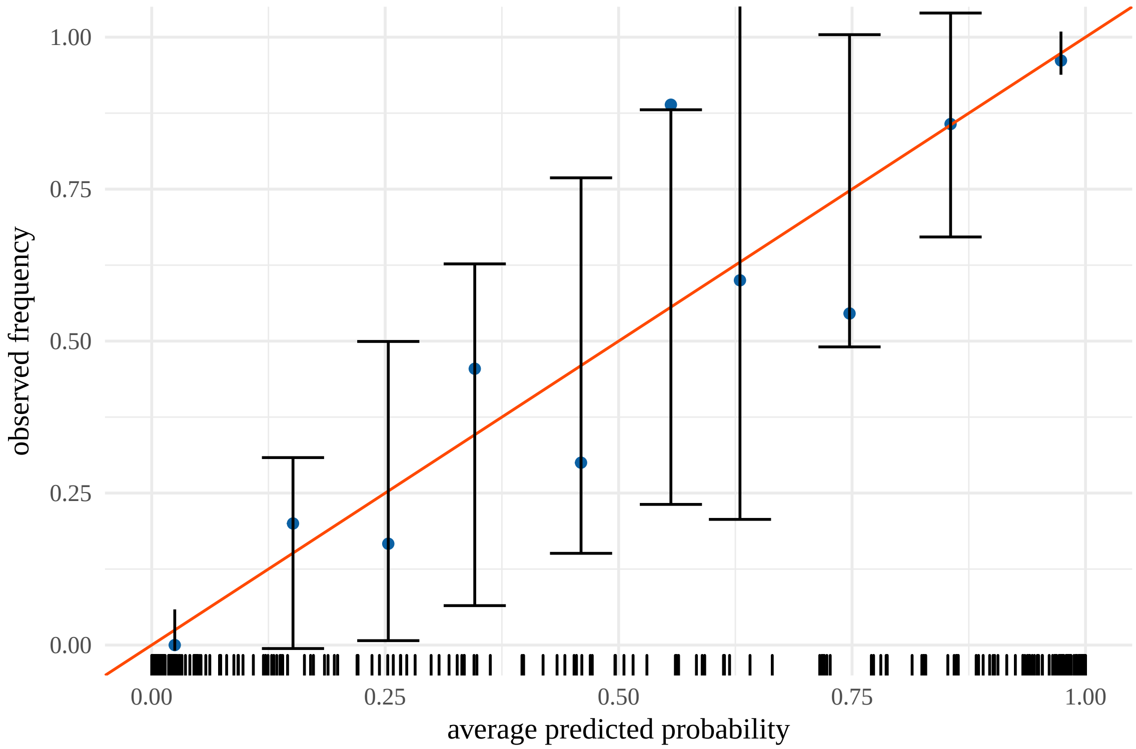
- The only thing we care about is whether the treatment led to an increase in happiness. Since the coefficient on treatment from our best model is -1.979 with a 95% CI of (-4.06, 0.101) (approximate), we suggest that the treatment probably led to a **decrease** in happiness. That is, people who were made to have more sex were less happy. This at first seems unexpected, but probably shouldn't be. People who are made to do something more or less than they had already decided to do are probably worse off. These couples chose to have sex with some particular frequency, and forcing them to double it is undesirable, just as forcing them to half it would likely be undesirable.

2 Writing functions

Below is one way to fix up these functions.

```
logit <- function(z){
  stopifnot(z>0, z<1) # need z between zero and 1
  log(z/(1-z))
}
ilogit <- function(z){
  stopifnot(is.finite(z)) # need finite z
  exp(z)/(1+exp(z))
}
sim.logistic <- function(n, beta.0, beta) {
  stopifnot((p <- length(beta)) >= 1, n > 0)
  x = matrix(rnorm(n*p), n, p)
  linear.parts <- beta.0+(x%*%beta)
  y <- rbinom(nrow(x), size=1, prob=ilogit(linear.parts))
  df = data.frame(y,x)
  if(ncol(x)==1) names(df)[2] = 'X1'
  return(df)
}
```

We simulate and test as instructed. Based on 250 replications, we derive a p-value of 0.796. We fail to reject the null hypothesis that the linear logistic regression is sufficient for describing these data. Finally, we produce a calibration plot and a confusion matrix.




```
##         prediction
## reality   0   1
##       0 119  13
##       1  14 104
```

The model appears to be reasonably well calibrated (though $n = 250$ is a bit small to use so many bins). In my case, the observed frequencies nearly always fall inside the confidence interval, with the exception of the bin with predicted probabilities around 0.55. From the confusion matrix, we appear to be reasonably accurate (a 0.108% error rate), and this accuracy is symmetric (we miss 0's and 1's at about the same rate).