Comparing methods

The `sexstudy.csv` dataset was collected by George Loewenstein, a behavioral sciences professor at Carnegie Mellon University. The study consisted of 65 married couples who were split into treatment and control groups. The treatment group was told to double the amount of sex they were having. Both groups completed a daily survey for a 3 month period. This data averages many of those measurements over the 3 month period. The goal is to see if more sex leads to more happiness. For more information see here.

- We're going to regress `happiness.scale.mean` (a measure of happiness, larger being happier) 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 will also use `age` and `yearsmar`.
- Do you need to convert the 0–1 variables to *factors* using, e.g. `factor(female)` in your `lm` call?
- 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.
- Why don't we need to smooth factors?
- The only thing we care about is whether the treatment led to an increase in happiness. Use the two estimated models to answer this question. What coefficient(s) should we look at?
- The following code extracts the leave-one-out CV from both models and prints it.

```
cvlin = mean(residuals(sslm)^2/(1-hatvalues(sslm))^2)
cvgam = mean(residuals(ssgam)^2/(1-ssgam$hat)^2)
c(cvlin,cvgam)
```

Which model do we prefer and why?

- Re-estimate the additive model but smoothing `age` and `yearsmar` together. Plot the partial response function(s).
- 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`. Do we still smooth `age2`? Plot the partial response function again.
- You have now made 3 more models. Modify the code to calculate their leave-one-out CV scores. Which model is better?
- Write a paragraph describing your conclusions. Does the treatment lead to more happiness? Use figures and numerics as necessary to support your conclusion.

Writing functions

In the chapter11 lecture, I gave you the following code to generate data from a logistic regression with 2 predictors (`x` variables):

```
logit <- function(z){ # can this take in any z?
  log(z/(1-z))
}
ilogit <- function(z){ # what about this one?
  exp(z)/(1+exp(z))
}
sim.logistic <- function(x, beta.0, beta, bind=FALSE) {
```

```
  linear.parts <- beta.0+(x%*%beta)
  y <- rbinom(nrow(x), size=1, prob=ilogit(linear.parts))
  if (bind) { return(cbind(x,y)) } else { return(y) }
}
```

1. Modify these functions in the following ways:

   - so that the `logit` and `ilogit` functions check that the arguments satisfy any necessary conditions.
   - so that `sim.logistic` doesn't require `x` as an input, rather it generates `x` within the function. Each entry of the `x` matrix should be normally distributed with mean zero and variance 1.
   - so that `sim.logistic` takes `n` (the number of observations) as an input
   - so that `beta` can be a vector of any length more than 1
   - so that `sim.logistic` returns a data frame (always) of `y` and `x`
   - so that `sim.logistic` checks that all the arguments satisfy any necessary conditions.

2. Use your modified function to generate `n=250` observations with `beta=c(3,2,1)` and `beta.0=0`. Then estimate:

   - a logistic regression model without intercept
   - a GAM without intercept, smoothing each component individually

3. Use the code from chapter11 (the deviance test) to test the GAM against the linear model. Is it safe to use the linear model, or do you need a GAM?

4. For whichever model you selected, produce a confusion matrix and make a calibration plot similar to the notes. Describe what these things tell you and discuss the quality of your chosen model.