# ST440/540 – Exam 2 - Due Monday, April 14

THIS IS AN EXAM - DO NOT DISCUSS THE PROBLEM WITH ANYONE (INCLUDING OTHER STUDENTS OR THE TA)! If you have questions, please email me.

The data you will analyze are (well, very similar to these data) from the paper Association of combined lead, cadmium, and mercury with systemic inflammation. The data has the following variables:

1. `CRP`: measure of systemic inflammation (outcome variable)

2. `cadmium, lead, mercury`: heavy metal concentrations in blood samples

3. `PFDeA, PFHxS, PFNA, nPFOA, sbPFOA`: PFAS concentrations in blood samples

4. `age, incomeration`: Subject descriptors

The outcome variable is `CRP` and all other variables are predictors. You can assume the subject descriptors have linear effects and you do not need to report their effects. The main interest is in the heavy metal and PFAS variables. The objective is to determine which, if any, of the heavy metals and PFAS concentrations are associated with `CRP` and whether the effects are linear or non-linear.

1. **Model description:** Describe your model (do not try the BKMR model they use in the paper, it is too complicated), including prior, and argue this is a reasonable approach for this task.

2. **Model comparisons:** Compare at least three models including some non-linear models and select a final model.

3. **Goodness of fit**: Verify that your choosen model fits the data well. If it does not fit well, select a more appropriate model and repeat steps 1–3.

4. **Variable importance**: Determine which predictors are the most important and summarize their effects

Your paper should be written as a professional document with clearly labeled figures and tables, full sentences and few spelling/grammatical errors. Include enough detail that the results could be replicated without looking through your code. Organize your report with subsections corresponding to the questions above. Summarize your analysis in a PDF document that is **no more than two pages long** (12 font, single space, standard margins, figues and tables count as part of the two pages); you will be penalized 10 points for each additional page. Append your code to the end of this document and submit a single document. All students should submit their exam on moodle by 1:30 PM on April 14.

HAVE FUN!

# Loading the data

```
> data   <- read.csv("E2.csv")
>
> lm.fit <- lm(log(CRP) ~ log(cadmium) + log(lead) +
+              log(mercury) + log(PFDeA) +
+              log(PFHxS) + log(PFNA) +
+              log(nPFOA) + log(sbPFOA) +
+              age + incomeratio, data=data)
> summary(lm.fit)


Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.364623   0.289878  -1.258  0.20856
log(cadmium)  0.046257   0.031931   1.449  0.14755
log(lead)    -0.120674   0.039708  -3.039  0.00240 **
log(mercury) -0.068870   0.027665  -2.489  0.01285 *
log(PFDeA)   -0.048814   0.043019  -1.135  0.25660
log(PFHxS)   -0.052986   0.036537  -1.450  0.14712
log(PFNA)    -0.056548   0.041656  -1.357  0.17474
log(nPFOA)   -0.076341   0.051844  -1.473  0.14100
log(sbPFOA)  -0.031222   0.103109  -0.303  0.76206
age           0.018076   0.001452  12.447  < 2e-16 ***
incomeratio  -0.042203   0.015191  -2.778  0.00551 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.207 on 2630 degrees of freedom
Multiple R-squared:  0.07437,   Adjusted R-squared:  0.07085
F-statistic: 21.13 on 10 and 2630 DF,  p-value: < 2.2e-16
```