# STA 35C: Homework 5

**Instructor: Akira Horiguchi**

Student name: ABCDE FGHIJ; Student ID: 123456789

Nov 5, 2025 (Wednesday), 22:59 PST

The assignment must be done in an R Markdown or Quarto document. The assignment must be submitted by the due date above by uploading:

- a .pdf file in GRADESCOPE (if you can knit/compile your .rmd to a .html file only, please save the created .html file as a .pdf file (by opening the .html file -> print -> save to .pdf)).

Email submissions will not be accepted.

Each answer has to be based on `R` code that shows how the result was obtained. The code has to answer the question or solve the task. For example, if you are asked to find the largest entry of a vector, the code has to return the largest element of the vector. If the code just prints all values of the vector, and you determine the largest element by hand, this will not be accepted as an answer. No points will be given for answers that are not based on `R`. This homework already contains chunks for your solution (you can also create additional chunks for each solution if needed, but it must be clear to which tasks your chunks belong).

There are many possible ways to write `R` code that is needed to answer the questions or do the tasks, but for some of the questions or tasks you might have to use something that has not been discussed during the lectures or the discussion sessions. You will have to come up with a solution on your own. Try to understand what you need to do to complete the task or to answer the question, feel free to search the Internet for possible solutions, and discuss possible solutions with other students. It is perfectly fine to ask what kind of an approach or a function other students use. However, you are not allowed to share your code or your answers with other students. Everyone has to write the code, do the tasks and answer the questions on their own.

During the discussion sessions, you may be asked to present and share your solutions.

```
set.seed(2025*4)  # do not change this; this helps to reproduce the "random" results
```

# 1. Cross-validation

We perform cross-validation on a simulated data set.

```
set.seed(1)
x <- runif(100) # 100 values being uniformly distributed (btw 0 and 1) are generated
y <- 1 + x -  x^2  + rnorm(100, 0, 0.1)
```

(a) Create a scatterplot where `y` is plotted against `x`. Describe your findings.

```
### Your Solution (Code)
```

(b) Use `lm()` to fit the three models below. Print the summary tables for the three fitted models and comment on your findings.

- Model I: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Model II: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$
- Model III: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

```
### Your Solution (Code)
```

(c) Calculate the leave-one-out-cross-validation mean squared error for each model I-III.

```
### Your Solution (Code)
```

(d) Calculate the $k$-fold cross-validation mean squared error for each model I-III for $k = 10$.

```
### Your Solution (Code)
```

(e) Which model has the smallest cross-validation error based on your results in (c) and (d)? Briefly explain why.

(f) Explain the individual concepts and the relationship between the validation set approach, leave-one-out cross-validation and $k$-fold cross-validation in about 1/2 page (maximum one page).

## 2. Ridge regression

ISLR2 Chapter 6, conceptual problem 4