STA 141A: Homework 5

Instructor: Akira Horiguchi

Student name: ABCDE FGHIJ; Student ID: 123456789

May 11, 2026 (Monday), 23:00 PDT

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(2026*2)  # do not change this; this helps to reproduce any "random" results
```

```
# can be helpful for debugging, but do not include output from this in your final submission
sessioninfo::session_info()
```

1. Cross-validation

Code examples from ISLR2 website

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. Linear Regression with a generated dataset

Code examples from ISLR2 website

```
GenerateData_v1 <- function(n) {
  x <- runif(n, min=-1, max=1)
  y <- -2 + 3 * x + rnorm(n, sd = 0.5)
  data.frame(x = x, y = y)
}
test_data <- GenerateData_v1(n=100000)
```

(a) Use the function `GenerateData_v1()` to generate a training dataset with $n = 10$ points. Fit a simple linear regression model to this dataset. What are the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$? Compute the training MSE. Compute the test MSE using the dataset `test_data`. How do the two MSEs compare?

- *Hint*: you may want to use the functions `lm()` and `predict()`.

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

(b) Do the same as in part (a), but now by generating a training dataset with $n = 100$ points.

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

(c) Do the same as in part (a), but now by generating a training dataset with $n = 1000$ points.

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

(d) As n increases, what values do the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ seem to converge to? As n increases, what value do the test MSEs seem to converge to, and how does this value relate to the noise standard deviation in the data generation process?

3. Linear Regression with the `anscombe` dataset

Consider the linear regression model

$$y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

and the `anscombe` dataset, whose columns have names `x1, x2, x3, x4, y1, y2, y3, y4`.

With this dataset, fit a linear regression model for each of the four x-y pairs. For each pair, do/answer the following:

1. What are the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$?
2. How about the R^2?
3. Plot y against x, along with the fitted regression line. (You can use `plot()` and `abline()` in base R, or `geom_line()` and `geom_smooth(method = "lm")` in ggplot2) For the plots, comment on your findings.

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

4. Linear Regression with the `mpg` data set

(a) Load the `mpg` dataset from `library(ggplot2)`. Add new variables as follows, and show the frequency table of those new variables.

- A binary variable of the car types: value `M` if the car type is either 2seater, compact, or midsize, and value `L` if the car type is either minivan, pickup, subcompact, or suv
- A binary variable of the type of transmission: value `auto` for auto cars, and value `manual` for manual cars

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

(b) Fit a linear regression model using `cty` as the response. For the predictors, use the variables defined in (a), as well as `displ`, `year`, and `cyl`. Comment on your findings.

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

(c) Plot the residuals against the fitted values. What can you conclude from this?

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