Task chapter 5 – resampling: Cross-validation

Perform leave-one-out cross validation (LOOCV) on a simulated data set.

1. **Generate a simulated data set**

```
set.seed(100)

x <- rnorm(100)

y <- x - 2*x^2 + rnorm(100)
```

What is n and what is p in this data set? Write the last line (`y <- ...`) as a mathematical equation for linear regression. What does `rnorm(100)` represent in this equation? Remember that you can find out what the functions `rnorm` and `set.seed` do by typing `?rnorm` in the R console (or googling).

2. **Data scatterplot** Produce a scatterplot of X against Y to get a visual impression of the data you created.

3. **LLS regression** Set a random seed and compute LOOCV errors that result from fitting four models using least squares:

```
Y = b0 + b1*X + epsilon

Y = b0 + b1*X + b2*X^2 + epsilon

Y = b0 + b1*X + b2*X^2 + b3*X^3 + epsilon

Y = b0 + b1*X + b2*X^2 + b3*X^3 + b4*X^4 + epsilon
```

Use `data.frame()` to create one data set that contains X *and* Y. Use `glm()` with the option `family = "gaussian"` to fit the linear model instead of `lm()`. The reason is that its results can be used with `cv.glm()` to compute cross validation. `cv.glm` is contained in the `boot` package (which you must install if not already on your system and then load). For LOOCV you only need to specify your data set and the variable to which you assigned your glm-fitted model as parameters for the `cv.glm` function (cf. also section 5.3.2 in James2021).

Careful! Formulas in R do not evaluate their contents. For example, in `y ~ x + x^2`, R would interpret the second term as a duplicate of x and drop it. You need to use the "as-is operator" `I()`: `y ~ x + I(x^2)`, which tells R to compute the values of `x^2` before attempting to use the formula.

4. **Interpret LOOCV errors** Which of the 4 models has the smallest LOOCV error? Plot all 4 models (resulting from the seed in subtask 3) on top of the 100 data points to explain the results.

5. **Regenerate data set and regressions** Repeat subtasks 1 - 3 but with a different random seed. Do the results differ? Why/not?