Prof. Yiping Lu
IEMS 304: Statistical Learning for Data Analysis
May 19, 2025

Homework 5

This homework is to give a brief reminder of R, RStudio, and statistical topics covered in IEMS 303.

Note: The homework is scored out of 100 points. The problems add up to 90 points, while the remaining ten points will be graded according to a writing rubric, given at the end of the assignment.

R/RStudio installation If you have not installed R and RStudio, follow the installation instructions outlined in https://posit.co/download/rstudio-desktop/. You are strongly encouraged to use R Markdown to integrate text, code, images and mathematics or you can you use the latex code we provide.

Question 1. Consider the `NCI60` data set in the `ISLR` package. I messed with the response variable (so the example isn't very realistic). Pretend the goal is to predict the severity of disease with the 6830 gene expression vectors.

```
install.packages(c("ISLR","glmnet"))
require(ISLR); require(glmnet)
data("NCI60")
NCI.genes <- NCI60$data
set.seed(47)
NCI.severity <- rnorm(length(NCI60$labs), mean=as.numeric(as.factor(
   NCI60$labs)),sd=2)
```

(a) Try running OLS on the data. What happens? Look at the lm output (including the coefficients) and explain. Try plotting or numerically summarizing your coefficients to say something meaningful. Please do not print out lists of 6830 things!!! Make your output meaningful!

(b) Can you use forward stepwise methods? Which gene do you think would be the first gene into the model? Note: you don't have to run forward (it's kind of a pain to write the code), just think about how you might assess which gene is most significant in a simple bivariate way.

(c) Can you go backwards? Explain.

(d) Run ridge regression on the data.
   i. Find the λ value that minimizes the cross validated error.
   ii. Provide plots for both the cross validated error and the coefficients as functions of λ
   iii. Did the variable(s) that you found in (b) have high coefficients? Note that the coefficient vector is 6831 long, because there is an intercept coefficient!

(e) Run lasso on the data.
   i. Find the λ value that minimizes the cross validated error.
   ii. Provide plots for both the cross validated error and the coefficients as functions of λ.
   iii. Did the variable(s) that you found in (b) have high coefficients? Note that the coefficient vector is 6831 long, because there is an intercept coefficient!

(f) Make a pairs plot of the three sets of coefficients (all, RR, Lasso). Give some interpretation of the plot. (n.b. You should remove the intercept and then cbind the non-intercept coefficients. Also, you may need as.numeric on the ridge regression and lasso coefficient vectors.)

Question 2. In this problem we are working on linear regression with regularization on points in a 2-D space. Figure plots linear regression results on the basis of three data points, $(0, 1)$, $(1, 1)$ and $(2, 2)$, with different regularization penalties.

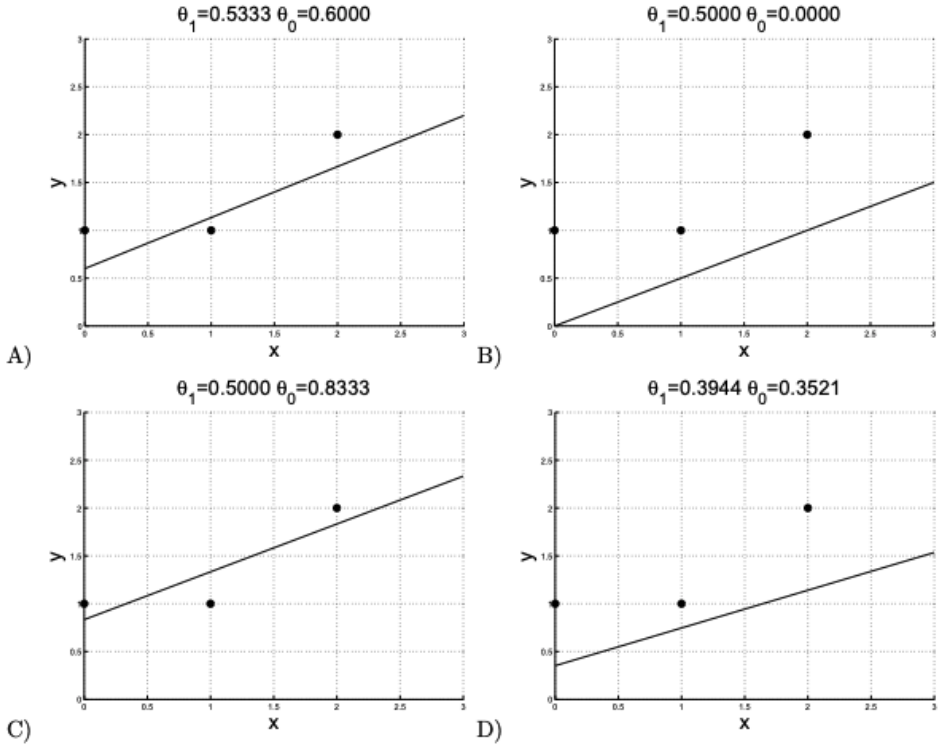

Figure 2: Plots of linear regression results with various regularization

As we all know, solving a linear regression problem amounts to solving the minimization

$$\arg\min_{\theta_0,\theta_1} \sum_{i=1}^{n} \Big(y_i - \theta_1 x_i - \theta_0\Big)^2 \; + \; R(\theta_0, \theta_1),$$

where R represents a regularization penalty which could be ℓ_1 or ℓ_2. In this problem, $n = 3$, $(x_1, y_1) = (0, 1)$, $(x_2, y_2) = (1, 1)$, and $(x_3, y_3) = (2, 2)$. We consider

$$R(\theta_0, \theta_1) = \lambda\Big(|\theta_1| + |\theta_0|\Big) \quad \text{or} \quad R(\theta_0, \theta_1) = \lambda\Big(\theta_1^2 + \theta_0^2\Big).$$

However, instead of computing derivatives to find the minimum, we adopt a geometric approach. Rather than letting the squared-error term and the regularization-penalty term vary simultaneously as functions of θ_0 and θ_1, we fix one and let the other vary. By imposing an upper bound r on the penalty, we replace the penalty term by the constraint $R(\theta_0, \theta_1) \; \leq \; r$ and solve the constrained problem

$$\min_{\substack{\theta_0,\theta_1 \\ R(\theta_0,\theta_1)\leq r}} \sum_{i=1}^{n} \Big(y_i - \theta_1 x_i - \theta_0\Big)^2.$$

The constrained formulation picks the smallest squared-error contour tangent to the level set $R(\theta_0, \theta_1) = r$. The tangent point yields the optimal (θ_0, θ_1).

By varying r and then minimizing over it, one recovers the full regularized solution:

$$\min_{\theta_0,\theta_1} \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2 + R(\theta_0, \theta_1) \; = \; \min_{r\geq 0} \Big\{ \min_{R(\theta_0,\theta_1)\leq r} \sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2 \Big\} + r.$$

Please assign each plot in Figure to one (and only one) of the following regularization methods. Use A, B, C or D.

(a) No regularization ($\lambda = 0$): $\sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2$.

(b) Ridge regression (ℓ_2 penalty) and λ being 5: $\sum_{i=1}^{n} (y_i - \theta_1 x_i - \theta_0)^2 + \lambda(\theta_1^2 + \theta_0^2)$, where $\lambda = 5$.

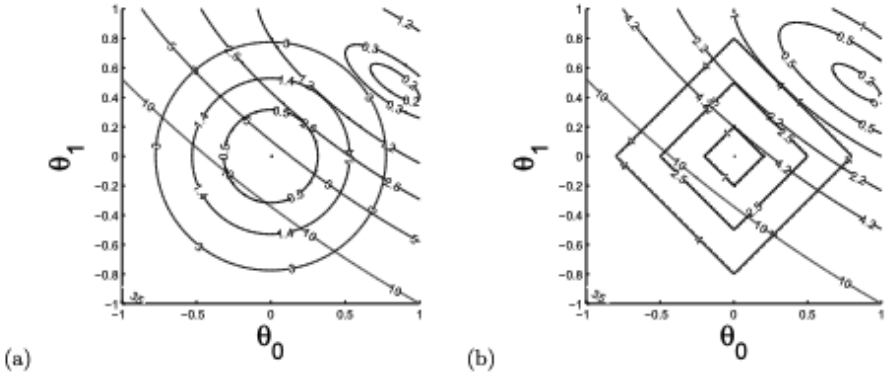

FIGURE 1. Contour plots of the decomposition for the linear regression problem with (a) L-2 regularization or (b) L-1 regularization where the ellipsis correspond to the square error term, and circles/squares correspond to the regularization penalty term.

(c) Lasso regression (ℓ_1 penalty) and λ being 5: $\sum_{i=1}^{n}(y_i - \theta_1 x_i - \theta_0)^2 + \lambda\Big(|\theta_1| + |\theta_0|\Big)$, where $\lambda = 5$.

(d) Ridge regression (ℓ_2 penalty) and λ being 1: $\sum_{i=1}^{n}(y_i - \theta_1 x_i - \theta_0)^2 + \lambda(\theta_1^2 + \theta_0^2)$, where $\lambda = 1$.

Hint: Look at the Counter Plot provided.

Rubric (10).

- The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow.
- Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels.
- All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.
- Code is either properly integrated with a tool like R Markdown or included as a separate R file. In the former case, both the knitted and the source file are included. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names.
- All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output. For full credit, all code runs, and the Markdown file knits (if applicable).