Econ 5253 - Spring 2025 **Problem Set 9** **Due: Apr. 8 beginning of class**

This problem set will give you practice in using cross-validation to tune linear regression prediction models via LASSO, ridge regression, and elastic net. The objective function of these models is:

$$\min_{\beta,\lambda,\alpha} \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda\left(\alpha\sum_k |\beta_k| + (1-\alpha)\sum_k \beta_k^2\right)$$

where $\hat{y}_i = x_i'\beta$ and where λ and α are parameters that must be *tuned* using cross-validation techniques. When $\alpha = 0$ we have the **ridge regression model**. When $\alpha = 1$ we have the **LASSO model**. When $\alpha \in (0,1)$ we have the **elastic net model**. Importantly, α must be in the closed interval $[0,1]$. As with the previous problem sets, you will submit this problem set by pushing the document to *your* (private) fork of the class repository. You will put this and all other problem sets in the path `/DScourseS25/ProblemSets/PS9/` and name the file `PS9_LastName.*`. Your OSCER home directory and GitHub repository should be perfectly in sync, such that I should be able to find these materials by looking in either place. Your directory should contain at least three files:

- `PS9_LastName.R` (you can also do this in Python or Julia if you prefer, but I think it will be much more difficult to use either of those alternative for this problem set)
- `PS9_LastName.tex`
- `PS9_LastName.pdf`

1. Type `git pull origin master` from your OSCER `DScourseS25` folder to make sure your OSCER folder is synchronized with your GitHub repository.
2. Synchronize your fork with the class repository by doing a `git pull upstream master`.
3. Install the following machine learning packages if you haven't already:
   - `tidymodels`
   - `glmnet`
4. Load the housing data from UCI, following the example in the lecture notes (Lecture 20).
5. Set the seed to `123456`.
6. Create two data sets called `housing_train` and `housing_test` using the `initial_split()` function from the `rsample` package, following the example in the lecture notes.

7. Create a new `recipe()` that takes the log of the housing value, converts `chas` to a factor, creates 6th degree polynomials of each of the continuous features (i.e. everything except `chas`), and linear interactions of each. To do so, add the following code to your script. What is the dimension of your training data (`housing_train`)? How many more X variables do you have than in the original `housing` data?

```
housing_recipe <- recipe(medv ~ ., data = housing) %>%
                  # convert outcome variable to logs
                  step_log(all_outcomes()) %>%
                  # convert 0/1 chas to a factor
                  step_bin2factor(chas) %>%
                  # create interaction term between crime and nox
                  step_interact(terms = ~ crim:zn:indus:rm:age:rad:tax:
                     ptratio:b:lstat:dis:nox) %>%
                  # create square terms of some continuous variables
                  step_poly(crim,zn,indus,rm,age,rad,tax,ptratio,b,
                     lstat,dis,nox, degree=6)
# Run the recipe
housing_prep          <- housing_recipe %>% prep(housing_train, retain
   = TRUE)
housing_train_prepped <- housing_prep %>% juice
housing_test_prepped  <- housing_prep %>% bake(new_data = housing_test)
# create x and y training and test data
housing_train_x <- housing_train_prepped %>% select(-medv)
housing_test_x  <- housing_test_prepped  %>% select(-medv)
housing_train_y <- housing_train_prepped %>% select( medv)
housing_test_y  <- housing_test_prepped  %>% select( medv)
```

8. Following the example from the lecture notes, estimate a LASSO model to predict log median house value, where the penalty parameter λ is tuned by 6-fold cross validation. What is the optimal value of λ? What is the in-sample RMSE? What is the out-of-sample RMSE (i.e. the RMSE in the test data)?

9. Repeat the previous question, but now estimate a ridge regression model where again the penalty parameter λ is tuned by 6-fold CV. What is the optimal value of λ now? What is the out-of-sample RMSE (i.e. the RMSE in the test data)?

10. In your .tex file, answer the questions posed in the preceding questions. Would you be able to estimate a simple linear regression model on a data set that had more columns than rows? Using the RMSE values of each of the tuned models in the previous two questions, comment on where your model stands in terms of the bias-variance trade-off.

11. Compile your .tex file, download the PDF and .tex file, and transfer it to your cloned repository on OSCER. There are many ways to do this; you may ask an AI chatbot or simply drag-and-drop using VS Code. Do **not** put these files in your fork on your personal laptop; otherwise git will detect a merge conflict and that will be a painful process to resolve.

12. You should turn in the following files: .tex, .pdf, and any additional scripts (e.g. .R, .py, or .jl) required to reproduce your work. Make sure that these files each have the correct naming convention (see top of this problem set for directions) and are located in the correct directory (i.e. `~/DScourseS25/ProblemSets/PS9`).

13. Synchronize your local git repository (in your OSCER home directory) with your GitHub fork by using the commands in Problem Set 2 (i.e. `git add`, `git commit -m "message"`, and `git push origin master`). More simply, you may also just go to your fork on GitHub and click the button that says "Fetch upstream." Then make sure to pull any changes to your local copy of the fork. Once you have done this, issue a `git pull` from the location of your other local git repository (e.g. on your personal computer). Verify that the PS9 files appear in the appropriate place in your other local repository.