Take-Home Exam – Applied Statistics
Start time: Wednesday (December 6) at Noon
End Time: Monday (December 11) at Noon

You are allowed to use any textbook, class notes, or other print or electronic reference materials that you would like, however you are not allowed to consult or collaborate with any other person, when writing your exam solution.

Columbia's policy on academic integrity stats that, as students of this class, you must be responsible for the full citations of others' ideas in all of your research papers and projects; you must be scrupulously honest when taking your examinations; you must always submit your own work and not that of another student, scholar, or internet agent.

Any breach of this intellectual responsibility is a breach of faith with the rest of our academic community. It undermines our shared intellectual culture, and it cannot be tolerated.

Please submit the following two files to `an3145@columbia.edu` by the deadline:

1) A PDF document with answers to the questions

2) An rmd file or python notebook file with necessary comments that can run on its own to reproduce the results.

The exam is based on the dataset `diabetes` from R package `lars`. The dataset has three matrices x, x2 and y. While x has a smaller set of independent variables, x2 contains the full set with quadratic and interaction terms. y is the dependent variable which is a quantitative measure of the progression of diabetes. We shall use x and y only.

Hypothesis Testing Based Approach Use cutoff 0.05 for p-values.

(a) Do forward selection and report the summary of the selected model.

(b) Do backward elimination and report the summary of the selected model.

Best Subset Selection

(c) Perform the best subset selection to identify which variables in the model are most important. Report the best model of each size.

(d) Construct a plot showing AIC, BIC, GCV, LOOCV and adjusted R^2 on the y-axis and model size on the x-axis. Is the curve monotonic? Explain. What model size minimizes each of the four criterion?

(e) Compare the models selected by these criteria. Why are they similar or different?

(f) Apply residual boostrap 100 times and report the frequency of the size of the selected model by each criterion. Summarize your observations.

Lasso

(g) Simulate a "noise" variable `age1` by randomly permute the observations for `age`. Similarly produce a "noisy copy" for each of the 10 covariates.

(h) Run Lasso with all 20 covariates, both the original 10 and the 10 added in (g). Plot the solution path.

(i) Perform 10 fold cross validation and identify the Lasso estimate. How many "noisy" variables were selected?

(j) Repeat (g)-(i) without plotting the solution path for 100 times, and report the frequency that each variable is selected in the final model. Summarize your observations.

OLS

(k) Run least squares estimate of y vs x, without variable selection. Run diagnosis for the fit to check potential violation of the the usual assumptions for linear regression,

(l) Construct confidence intervals for the regression coefficients.

(m) Construct bootstrap, both case resampling and residual resampling, confidence intervals for the linear regression coefficients. Compare these confidence intervals with those obtained in (l).