The iod.csv file contains another large data set from two kidney disease studies (MDRD Modificiation of Diet in Renal Disease) and AASK African-American Study of Kidney Disease). The variables included are the following:

WEIGHT (kg)

BMI body mass index

GFR glomerular filtration rate

UCRE urine creatinine

UUN urine urea nitrogen

UPHO urine phosphorus

SUN serum urea nitrogen

SCR serum creatinine

TCHOL total cholesterol

ALB albumin

HBA1C hemoglobin AIC

PHOS serum phosphorus

TRIG triglycerides

LDL low density lipoprotein (cholesterol)

HDL high density lipoprotein (cholesterol)

HB hemoglobin

MAP mean arterial pressure

UPRO urine protein

BSA body surface area

SODIUM sodium

GLUC glucose

BLACK black race

HEIGHT height (cm)

AGE age

FEMALE female

CYS serum cystatin

DBP diastolic blood pressure

SBP systolic blood pressure

CRP C-reactive protein

DIAB diabetes

HBPSTATUS high blood pressure (1/0)

Use GFR as your outcome and construct a predictive model for it using a) stepwise regression; b) ridge regression; c) lasso regression using cross-validation to choose your model form. For this exercise, just use the variables as they are given. In the next homework, you will consider transformations and nonlinear functions.

Describe your findings in clearly written text, tables and figures discussing both the differences between the model findings and consistencies. Which factors are predictive? How well do the models predict the outcomes? Consider how you can get a good estimate of test error using cross-validation. At the end, refit using the best fitting model of each type on the whole dataset.

Biostat PhD problems (Extra Credit for everyone else):

1. The ridge estimator can be written

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

and the lasso minimizes

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Show that this minimization using penalty terms is equivalent to the following constrained optimizations

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} s.t. \sum_{j=1}^{p} \beta_j^2 \leq s$$

and

$$\min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} s.t. \sum_{j=1}^{p}|\beta_j|\leq s$$

.

State the relationship between s and λ.

2. Show that the ridge solution can also be written as

$$\hat{\beta}^{ridge}=\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

where $\mathbf{I}$ is the identity matrix.

3. Show that the ridge regression estimate is the mean (and mode) of the posterior distribution under a Gaussian prior $\beta\sim N\left(0,\tau^2\mathbf{I}\right)$ and Gaussian sampling model $\mathbf{y}\sim N\left(\mathbf{X}\beta,\sigma^2\mathbf{I}\right)$. Find the relationship between λ and the variances τ^2 and σ^2. [Hint: Note the form of the log posterior density of β].

4. Show that C_p and AIC are equivalent for a linear model with Gaussian errors

5. Prove that the probability that an observation appears in a specific bootstrap sample is approximately 0.632. (Hint: Calculate the probability that the first (or any other) bootstrap observation is a specific observation (e.g., the jth) from the original dataset]

6. Augment the centered design matrix $\mathbf{X}$ with p additional rows $\sqrt{\lambda}\mathbf{i}$ and augment $\mathbf{y}$ with p zeros. Show that the least squares solution to this augmented data set is the same as the ridge solution to the original dataset. Therefore, by introducing artificial data with response 0, the least squares coefficients are shrunk toward zero.