STAT 540, Fall 2015

Due on **12:00 A.M., Monday, Nov 30th**

Homework 11

Homework format for all STAT 540 homework this term: Please label all problems clearly and turn in an organized homework assignment. You don't need to spend hours producing beautifully typeset homework, but you won't get credit if we can't find or read your answer. Unless noted otherwise, turn in the following (as appropriate for the problem).

- Theoretical derivation (when asked for).
- Numerical results **with an explanation of your solution**, written in complete sentences. If computer code is absolutely necessary to provide context here, then include it–nicely formatted–within the solution (otherwise, see below).
- Appropriate graphics. Use informative labels, including titles and axis labels. Try to put multiple plots on the page by using, for example, the R command `par(mfrow=c(2,2))`.
- **Only as necessary**: Final clean computer code used to answer the problem **attached to the end of your homework**. Only include the rare code excerpts without which we wouldn't be able to figure out what you did. Annotate your code. Number and order the code in order of the problems. When in doubt, leave it out; consider that we will probably never read it.
- Some problems will be relatively open-ended, such as "Here are some data. Analyze them and write a report." I will provide further instructions about reports later. They should be self-contained, with suitable EDA, graphs, numerical results, and **scientific interpretation**. No computer code should be included. The report should be concise: "no longer than necessary".

(1) Model selection (practice): The `modelsel1.txt` data set contains four predictor variables and $n = 50$ observations. Using the `R` package `leaps`, you can easily get Mallows' C_p, AIC, BIC, etc.

(a) Find the best model using the stepwise selection, using $\alpha_{\text{entry}} = \alpha_{\text{remove}} = 0.2$.

(b) Do you end up with the same model using using $\alpha_{\text{entry}} = \alpha_{\text{remove}} = 0.1$?

(c) Use `R` to calculate Mallow's C_p for all possible models. Identify the variables in the 3 models with the three smallest C_p values.

(d) If you use AIC or BIC, do you select the same best model? The same best three models?

(2) Model building examples – For each of the following, give: an appropriate regression model and define the X variables in your model (including indicator variables).

Some examples are looking for estimates. For these: Indicate how the desired quantities could be estimated from the regression parameters. You do not need to worry about standard errors or inference. Other parts are looking for a test. For these: Indicate how you would construct that test. Your answer could be "a t-test of (indicate a regression parameter or linear combination of regression parameters) = 0 (or other value)". It could be "an F test comparing (indicate a pair of models)" or it could be something else. I do not need formula for the test statistics.

(a) A study is comparing the energy content of constant-sized pieces of firewood from different tree species. If you are burning wood to heat a room or a house, a higher energy content is a good thing. One complication is that the energy released depends on the moisture content of the firewood, which is hard to standardize. You have studied three species (Red Oak, White Pine and Black Walnut). You believe that the relationship between energy content and moisture is linear with the same slope for each species. You want to estimate the difference in energy content at 10% moisture content between White Pine and Red Oak.

(b) Same study as above, except now you wish to test the null hypothesis that the three species have the same energy content at 10% moisture. Again, assume that all three species have the same slope.

(c) Same study as above, except that now you assume that the three species have different slopes (for the association of moisture content on energy). You want to estimate the difference in energy content at 10% moisture between White Pine and Red Oak.

(d) Assume that athletic performance for males in a certain sport can be described by a quadratic function of age. You wish to estimate the age at which performance is maximum.

(e) Assume that athletic performance for males and females in a certain sport can be described by quadratic functions of age. You are willing to assume that the curvature (β_2) is the same for both. You wish to test whether the age of maximum performance is the same for males and females. Hint: The intercepts β_0 are probably not the same for males and females.

(f) Toxicologists study the effect of chemical contaminants. They often summarize their data by the quantity EC_{50}, the concentration of chemical that leads to a 50% reduction in response from the control response (at dose= 0). Assume that the relationship between Y, a measure of effect, is linearly related to X, the dose of a particular chemical. The intercept, β_0, is the expected response at dose= 0. You wish to estimate EC_{50}.

(g) Education researchers are studying whether watching television impacts the performance of graduate students. For each student in a class, they have Y, the exam score, and X, the number of hours spent watching television during the week prior to the exam. They assume that the relationship between Y and X is linear up to 20 hours.

After $X = 20$ hours, there is no relationship, i.e. the slope is 0 for $X > 20$. They wish to estimate the slope from 0 to 20 hours and the expected difference between light television watching (3 hours) and heavy watching (25 hours).

(3) A rehabilitation center researcher was interested in examining the relationship between physical fitness of patients undergoing corrective knee surgery and time required in physical therapy until successful rehabilitation. A sample of 24 patient records was randomly selected from the male patients ranging in age from 18 to 30 years who had undergone corrective knee surgery during the past year. Each patients was classified into one of three physical fitness categories (below average, average, above average) corresponding to their physical fitness prior to the corrective knee surgery. The number of days required for successful completion of physical therapy (Y) was recorded for each patient along with the age (X) of the patient. The data are shown below:

Below Average		Average		Above Average	
Age (X)	Rehab Time (Y)	Age (X)	Rehab Time (Y)	Age (X)	Rehab Time (Y)
18.3	29	20.8	30	22.7	26
30.0	42	25.2	35	28.7	32
26.5	38	29.2	39	18.9	21
28.1	40	20.0	28	18.0	20
29.7	43	21.5	31	21.7	23
27.8	40	22.1	31	20.0	22
19.8	30	19.7	29		
29.3	42	24.7	35		
		20.2	29		
		22.9	33		

(a) First consider the three physical fitness categories as the levels of the physical fitness factor and **REPORT** the one-way ANOVA for the rehabilitation times.

(i) Does the F-test indicate any differences among mean rehabilitation times for the three physical fitness categories? Explain.

(ii) Use the Bonferroni method (with an experimentwise type I error level no larger than 0.05) to compare the mean rehabilitation times for all three pairs of physical fitness categories. State your conclusions.

(iii) Construct a 95% confidence interval for the difference between the mean rehabilitation times for the above average and below average physical fitness categories. What does this interval indicate?

(b) Redo the analysis in part (a) using age as a covariate. Use a model for which the slope on age is the same for each physical fitness category (parallel lines). **REPORT** an

ANOVA table using Type I sums of squares with effects of pre-surgery physical fitness adjusted for age.

(i) Does the F-test indicate any differences among mean rehabilitation times for the three physical fitness categories when rehabilitation times are adjusted for age differences of the patients? Explain.

(ii) Use the Bonferroni method (with an experimentwise type I error level no larger than 0.05) to compare the age adjusted mean rehabilitation times for all three pairs of physical fitness categories. State your conclusions.

(iii) Construct a 95% percent confidence interval for the difference between the mean rehabilitation times for the above average and below average physical fitness categories after adjusting for the ages of the patients. Compare the length of this interval to the length of the interval from part (a-iii). Did adjusting for age improve the estimation of the difference in mean rehabilitation times for these two physical fitness categories?

(iv) Plot the rehabilitation times against the ages of the patients, using a different symbol for each physical fitness category, and insert the estimated regression lines. Does it appear that the model relating rehabilitation time to patient age with parallel straight lines is appropriate for these data? Explain.

(v) For each physical fitness category, plot the residuals versus the patient age. What do these plots indicate?

(vi) Construct a normal probability plot from the residuals. What can you conclude from this analysis?

(c) Now fit a model with a different slope on age for each of the physical fitness categories. By comparing the results to those for part (b), test the null hypothesis that the regression lines have the same slope. Report the p-value for the test and state you conclusion.

(d) The test performed in part (c), that compares the models from parts (c) and (b), is reasonable only if the model in part (c) is appropriate for these data, i.e., there is a straight line relationship between rehabilitation time and patient age with each of the physical fitness categories, the random errors are normally distributed with homogeneous variances across physical fitness categories. Does an examination of residuals from fitting the model in part (c) indicate any serious violations of these model assumptions?

(e) Regardless of what you discover in parts (c) and (d), use the parallel lines model from part (b) to estimate the mean number of days of therapy required for 30 year old male patients of average physical fitness. Report the standard error for your estimate and a 99% confidence interval.

(f) Use the parallel lines model from part (b) to construct a 99% percent prediction interval for the actual number of days of therapy required for 30 year old male patients of average physical fitness.

(4) Overfitting and underfitting problem (page 9 on lecture 13) Variable selection is important for regression since there are problems in either using too many irrelevant or too few (omitted) variables in a regression model. Consider the linear regression model $y_i = \mathbf{x}_i'\boldsymbol{\beta}+\epsilon_i$ where $\mathbf{x}_i \in \mathbb{R}^p$ and the errors are i.i.d. satisfying $\mathbb{E}(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$. Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be the response vector and $\mathbf{X}$ be the $n \times p$ design matrix. Assume that only the first p_0 variables are important, let $A = \{1, \ldots, p\}$ be the index set of the full model and $A_0 = \{1, \ldots, p_0\}$ be the index set for the true model. The true regression coefficients are $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_{A_0}^{*'}, \mathbf{0}')'$. Now consider three different modeling strategies:

- S1: fit the full model, and denote the full design matrix as $\mathbf{X}_A$ and corresponding OLS estimator by $\widehat{\boldsymbol{\beta}}_A^{ols}$
- S2: fit the true model using the first p_0 predictors, and denote the design matrix by $\mathbf{X}_{A_0}$ and OLS estimator by $\widehat{\boldsymbol{\beta}}_{A_0}^{ols}$
- S3: fit a subset model using only the first q predictors with $q < p_0$, and denote the design matrix by $\mathbf{X}_{A_1}$ and OLS estimator by $\widehat{\boldsymbol{\beta}}_{A_1}^{ols}$

Answering the following two questions

(a) One possible consequence of including irrelevant variables is that the prediction are not efficient, that is having large variance though they are unbiased. For any $\mathbf{x} \in \mathbb{R}^p$, show that

$$\mathbb{E}(\mathbf{x}_A'\widehat{\boldsymbol{\beta}}_A^{ols}) = \mathbf{x}_{A_0}'\boldsymbol{\beta}_{A_0}^*$$

and

$$\mathrm{Var}(\mathbf{x}_A'\widehat{\boldsymbol{\beta}}_A^{ols}) \geq \mathrm{Var}(\mathbf{x}_{A_0}'\widehat{\boldsymbol{\beta}}_{A_0}^{ols})$$

where $\mathbf{x}_{A_0}$ consists of the first p_0 elements of $\mathbf{x}$.

(b) One consequence of excluding important variables in a linear model is that the prediction are biased, though they admit smaller variances. For any $\mathbf{x} \in \mathbb{R}^p$, show that

$$\mathbb{E}(\mathbf{x}_{A_1}'\widehat{\boldsymbol{\beta}}_{A_1}^{ols}) \neq \mathbf{x}_{A_0}'\boldsymbol{\beta}_{A_0}^*$$

and

$$\mathrm{Var}(\mathbf{x}_{A_1}'\widehat{\boldsymbol{\beta}}_{A_1}^{ols}) \leq \mathrm{Var}(\mathbf{x}_{A_0}'\widehat{\boldsymbol{\beta}}_{A_0}^{ols})$$

where $\mathbf{x}_{A_1}$ consists of the first $q < p_0$ elements of $\mathbf{x}$.

(5) SSE for backward selections. As discussed on page 20 in lecture 13, backward selection starts with the model with all variables. At each step, it removes the variable making the smallest contribution. Assume that there are currently k variables in the model and the corresponding design matrix is $\mathbf{X}_1$. Verify that the expression on page 20 in lecture 13, i.e. the new sum of squared errors (SSE) from fitting resulting from deleting the jth ($1 \leq j \leq k$) variable from the current the k-variable model is

$$SSE_{k-1} = SSE_k + \frac{(\widehat{\beta}_{1j})^2}{s_{jj}}$$

where $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\beta}_{11}, \ldots, \widehat{\beta}_{1k})'$ is the vector of current regression coefficients and s_{jj} is the jth diagonal entries of $(\mathbf{X}_1'\mathbf{X}_1)^{-1}$.

(6) For model $y = f(\mathbf{x}) + \epsilon$ with $\text{var}(\epsilon) = \sigma^2$ and some function f, assume that the prediction is a linear smoother/fitting operator that $\widehat{\mathbf{y}} = \mathbf{S}\mathbf{y}$, show that

$$\sum_{i=1}^{n} \text{cov}(\widehat{y}_i, y_i) = \text{tr}(\mathbf{S})\sigma^2$$

which justifies the effective number of parameters in lecture.

(7) (Penalized least squares and orthogonal design) Consider the special case of an orthogonal design matrix $\mathbf{X}'\mathbf{X} = \mathbf{I}_n$ the penalized least squares problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} J(|\beta_j|)$$

becomes solving p one-dimensional shrinkage problems

$$\min_{\beta_j} (\beta_j - \widehat{\beta}_j^{ols})^2 + \lambda J(|\beta_j|).$$

Now consider four types of penalized least squares problems with orthogonal $\mathbf{X}$.

(a) Show that the ridge estimates are given by

$$\widehat{\beta}_j^{ridge} = (1 + \lambda)^{-1} \widehat{\beta}_j^{ols}$$

(b) The nonnegative garrot estimator is defined by

$$\min_{\mathbf{c}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} c_j x_{ij} \widehat{\beta}_j^{ols})^2 + \lambda \sum_{j=1}^{p} c_j, \text{ subject to } c_j \geq 0.$$

When $\mathbf{X}$ is orthogonal, the nonnegative garrote estimator seeks a set of nonnegative scaling factors c_j for $j = 1, \ldots, p$ by solving

$$\min_{c_j} (c_j \widehat{\beta}_j^{ols} - \widehat{\beta}_j^{ols})^2 + \lambda c_j \text{ such that } c_j \geq 0.$$

Show that the solution has the expression

$$\widehat{c}_j = [1 - \lambda/(2(\widehat{\beta}_j^{ols})^2)]_+$$

where $[u]_+ = \max(u, 0)$. Therefore, the final NG estimator for the $\boldsymbol{\beta}$ is

$$\widehat{\beta}_j^{ng} = [1 - \lambda/(2(\widehat{\beta}_j^{ols})^2)]_+ \widehat{\beta}_j^{ols}$$

(c) Show that the LASSO solution is given by

$$\widehat{\beta}_j^{lasso} = \text{sign}(\widehat{\beta}_j^{ols})[|\widehat{\beta}_j^{ols}| - 2\lambda]_+$$

(d) When $\mathbf{X}$ is orthogonal, the sparse constrained least square problems, i.e. the L_0 penalty or hard thresholding estimation, solve

$$\min_{\beta_j} (\beta_j - \widehat{\beta}_j^{ols})^2 + \lambda \mathbb{I}(|\beta_j \neq 0).$$

Show that the L_0 estimator is

$$\widehat{\beta}_j^0 = \text{sign}(\widehat{\beta}_j^{ols})\mathbb{I}(|\widehat{\beta}_j^{ols}| > \sqrt{\lambda})$$

(8) As discussed in lecture, Fan and Li (2001) proposed a penalized least squares using a smoothly clipped absolute deviation (SCAD) penalty of the form

$$q_\lambda(|w|) = \begin{cases} \lambda|w| & \text{if } |w| \leq \lambda \\ -\frac{(|w|^2 - 2a\lambda|w| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |w| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |w| > a\lambda \end{cases}$$

where $a > 2$ and $\lambda > 0$ are tuning parameters.

(a) Show that q_λ has a continuous first-order derivative everywhere except at the origin.

(b) Show that q_λ is not convex.

(c) Show that q_λ can be decomposed as the difference of two convex functions that

$$q_\lambda(|w|) = q_{\lambda,1}(|w|) - q_{\lambda,2}(|w|)$$

where q_1 and q_2 are convex and satisfy

$$q'_{\lambda,1}(|w|) = \lambda,\ q'_{\lambda,2}(|w|) = \lambda\left(1 - \frac{[a\lambda - |w|]_+}{(a-1)\lambda}\right)\mathbb{I}(|w| > \lambda)$$

(9) (Simulation studies) Consider $\boldsymbol{\beta} = (2.5, 3, 0, 0, 0, 1.5, 0, 0, 0, 4, 0)$ as the true coefficient so the correct number of important variable is 4. Generate covariates $\mathbf{x}$ using an autoregressive structure of order one, that is $x_j = \rho x_{j-1} + u_j$ with $u_j \sim N(0, 1)$. In this simulation study, you can specify $x_1 \sim N(0, 1)$. Consider $\rho = 0, 0.5, 0.9$, corresponding to no dependence, moderate dependence, and high dependence.

Consider model $y = \mathbf{x}'\boldsymbol{\beta} + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ with $\sigma = 1, 2, 3$. Let $n = 50$. The simulation study will be repeated for 500 times to get approximation of prediction errors. For each of the traditional and shrinkage methods for variable selection, five numerical summaries of the sampling distribution need to be computed.

- the average MSE,
- the number of explanatory variables correctly found to be zero is given (the true value is six),
- the number of explanatory variables incorrectly found to be zero (the correct value is zero; the worst value is four),
- the probability that the method selected the correct model; this is the fraction of times the correct model was chosen over the 500 iterations,
- the inclusion probabilities of each of the explanatory variables; that is the fraction of times the variable x_j being selected over the 500 iterations.

An example of data generation `R` code is listed below.

```
##INPUT parameters
N = 500 #number of simulations
```

```
n = 50 #sample size
p = 10 #total number of covariates
sigma = 3 #the standard deviation of noise
rho = 0.0 #the correlation coefficient in AR or CS
truebeta = c(2.5,3,0,0,0,1.5,0,0,4,0)

##OUTPUT results
betahat = matrix(0,N,p)
beta0hat = rep(0,N)
modelerr = rep(0,N)
varprob = matrix(0,N,p)
gVarprob = matrix(0,N,p)

##specify the AR(rho) covariance matrix for X
Xcov<-matrix(0,p,p)
for (i in 1:p)
{
    {for (j in 1:p)
        Xcov[i,j]<-rho(abs(i-j))
    }
}
svd.Xcov<-svd(Xcov)
v<-svd.Xcov$v
d<-svd.Xcov$d
D<-diag(sqrt(d))
S<-(v)%*%D%*%t(v)

## LOOP starts here
for (i in 1:N)
{
    set.seed(2009+i)
    ##generate X
    Ztr<-matrix(rnorm(n*p),n,p)
    X<-Ztr%*%S
    #generate y
    ymean<-X%*%truebeta
    y<-ymean+sigma*rnorm(n)

    # Here: your function or code for selecting models based on X and y...
    # output the results betahat, beta0hat,...
```

```
}
```

(a) Consider using AIC, BIC, GCV to select model first. Using the `R` package `leaps` for the computation, and report summary tables with some discussion on what you found. You should have six tables in total that two tables for each ρ: the first table should have the first four measurements compared across AIC, BIC, GCV for different σ, the second table should have the last numerical summary (for ten variables) compared across AIC, BIC, GCV for different σ. An example code is displayed below.

```
library(leaps)
forward_fit <- regsubsets(Xtr,ytr,method="forward")
aic <- which.min(2*(2:(p+2))+n*log(forward_fit$rss/n)+n+n*log(2*pi))
bic <- which.min(log(n)*(2:(p+2))+n*log(forward_fit$rss/n)+n+n*log(2*pi))
gcv <- which.min(forward_fit$rss/(n*(1-(2:(p+2))/n)2))
```

(b) Consider comparing shrinkage methods including elastic net (Enet), LASSO, Adaptive LASSO (ALASSO), and SCAD penalties. In this HW, you could chose a fixed λ although in practice λ is chosen using the BIC for all methods.

The packages `lars` and `elasticnet` in `R` can be used to solve elastic net, LASSO, ALASSO. Particularly, for ALASSO you need to specify the weights $\mathbf{w} = (w_1, \ldots, w_p)$ then transform the design matrix $\mathbf{X}$ to $\mathbf{X}^*$ of dimension $n \times p$ as $x^*_{ij} = x_{ij}/w_j$ so that $\widehat{\boldsymbol{\beta}}^{alasso} = \widehat{\boldsymbol{\beta}}^{lasso,*}/\mathbf{w}$ componentwisely, where $\widehat{\boldsymbol{\beta}}^{lasso,*}$ is the LASSO solution with $\mathbf{X}^*$. The very new package `ncvreg` can be used to solve either LASSO or SCAD.

```
> library(lars)
> lasso_fit <- lars(X,y,type="lasso")
> lasso_coef <- coef(lasso_fit, type="coef",mode="lambda")

> library(elasticnet)
> enet_fit <- enet(X,y,lambda)
> enet_coef <-coef(enet_fit, type="coef",mode="penalty")

> library(ncvreg)
> data(prostate)
> X <- as.matrix(prostate[,1:8])
> y <- prostate$lpsa
> fit_SCAD <- ncvreg(X,y,penalty="SCAD")
> plot(fit_SCAD,main=expression(paste("SCAD, ",gamma,"=",3)))
```

Report summary tables with some discussion on what you found like did for part (a).