Problem Set 7

1. Use the data in MROZ.raw to answer this question. The data consist of 753 observations. Before answering any of parts a.-b., remove 253 observations which will be used for an out-of-sample comparison in part c. Ideally, these would be the same observations left out when you answered problem 1 on Problem Set 5.

   a. Estimate the model $\mathrm{E}[\text{inlf}|X = \{\text{kidslt6, kidsge6, age, educ, repwage, faminc, exper}\}] = \mathcal{P}_K(X)' \beta$ by lasso with penalty parameter chosen by cross-validation. Carefully explain how you construct the dictionary of approximating functions $\mathcal{P}_K(X)$?

   b. Estimate the model $\mathrm{E}[\text{inlf}|X = \{\text{kidslt6, kidsge6, age, educ, repwage, faminc, exper}\}] = \Lambda(\mathcal{P}_K(X)' \beta)$ where $\Lambda(\cdot)$ is the logistic cdf by $\ell_1$-penalized logistic regression with penalty parameter chosen by cross-validation.

   c. Use the 253 observations you held out to compare the estimates obtained in parts a.-b. and the estimates obtained in problem 1 on Problem set 5. Calculate the mean square forecast error as $\frac{1}{253} \sum_{i \in \text{hold-out}} (\hat{g}_j(x_i) - y_i)^2$ and the misclassification rate $\frac{1}{253} \sum_{i \in \text{hold-out}} 1(\hat{y}_{j,i} \neq y_i)$ where $\hat{y}_{j,i}$ is the Bayes-classifier based on model $j$ - i.e. $\hat{y}_{j,i} = 1(\hat{g}_j(x_i) \geq .5)$, and $\hat{g}_j(x_i)$ are the fitted values obtained from each of the competing models. Which procedure performs best according to each metric? Do the performance discrepancies seem large? [Note: Assuming independent sampling, you can compute a standard error for the mean square forecast error and for the misclassification rate conditioning on the estimated model.]

2. Use the data in CreditCardDefault.xls to answer this question. The data consist of 30000 observations. Before answering any of parts a.-b., remove 10000 observations which will be used for an out-of-sample comparison in part c. Ideally, these would be the same observations left out when you answered problem 2 on Problem Set 5.
a. Estimate the model \( E[\text{default}|x_1, ..., x_{23}] = p_K(x_1, ..., x_{23})'\beta \) for \( x_1, ..., x_{23} \) defined in CreditCardDefault.des by lasso with penalty parameter chosen by cross-validation. Carefully explain how you construct the dictionary of approximating functions \( p_K(x_1, ..., x_{23}) \)?

b. Estimate the model \( E[\text{default}|x_1, ..., x_{23}] = \Lambda(p_K(x_1, ..., x_{23})'\beta) \) for \( x_1, ..., x_{23} \) defined in CreditCardDefault.des and \( \Lambda(\cdot) \) the logistic cdf by \( \ell_1 \)-penalized logistic regression with penalty parameter chosen by cross-validation.

c. Use the 10000 observations you held out to compare the estimates obtained in parts a.-b. and the estimates obtained in problem 2 on Problem set 5. Calculate the mean square forecast error as \( \frac{1}{10000} \sum_{i \in \text{hold-out}} (\hat{g}_j(x_i) - y_i)^2 \) and the misclassification rate \( \frac{1}{10000} \sum_{i \in \text{hold-out}} 1(\hat{y}_{j,i} \neq y_i) \) where \( \hat{y}_{j,i} \) is the Bayes-classifier based on model \( j \) - i.e. \( \hat{y}_{j,i} = 1(\hat{g}_j(x_i) \geq .5) \), and \( \hat{g}_j(x_i) \) are the fitted values obtained from each of the competing models. Which procedure performs best according to each metric? Do the performance discrepancies seem large? [Note: Assuming independent sampling, you can compute a standard error for the mean square forecast error and for the misclassification rate conditioning on the estimated model.]

3. [Post-selection inference example] Consider a linear regression model

\[
Y_i = \beta_1 X_{1,i} + \beta_{2,n} X_{2,i} + \varepsilon_i \\
X_{1,i} = \pi X_{1,i} + v_i
\]

where \((\varepsilon_i, v_i)' \sim N(0, \text{diag}(\sigma^2, \kappa^2))\) are iid across \( i \) and independent of the \( n \times 2 \) design matrix \( X \) and the second equation simply parameterizes the covariance between \( X_1 \) and \( X_2 \). Suppose that the parameter of interest is \( \beta_1 \) and that you are unsure of whether \( X_2 \) should be included in the model (i.e. you are unsure about whether \( \beta_{2,n} = 0 \)). Let \( \hat{\beta}_1 \) and \( \tilde{\beta}_1 \) be the conventional OLS estimators of \( \beta_1 \) and \( \beta_{2,n} \) obtained by regressing \( Y \) on \( X_1 \) and \( X_2 \), and let \( s_{\hat{\beta}_1} \) and \( s_{\tilde{\beta}_1} \) denote the corresponding standard error estimators. Let \( \hat{\beta}_1 \) be the conventional OLS estimator of \( \beta_1 \) obtained by regressing \( Y \) on \( X_1 \) (i.e. excluding \( X_2 \)), and let \( s_{\hat{\beta}_1} \) be the corresponding standard error estimator.

a. Under the assumption that \( \beta_{2,n} = 0 \), show that \( \tilde{\beta}_1 \) is consistent, asymptotically normal, and has variance less than or equal to that of \( \hat{\beta}_1 \).
b. Let $t_{\hat{\beta}_2} = \frac{\hat{\beta}_2 - \beta_2}{s_{\hat{\beta}_2}}$. Show that $\Pr(|t_{\hat{\beta}_2}| > c_n) \to 1$ for $c_n = \log(n)t_{n-2}^{-1}(0.975) \approx 2\log(n)$ when $\beta_{2,n} = \delta$ with $|\delta| > 0$ where $t_{n-2}$ denotes the cdf of $t_{n-2}$ random variable. Similarly show that $\Pr(|t_{\hat{\beta}_2}| \leq c_n) \to 1$ when $\beta_{2,n} = 0$.

c. Consider the estimator $\hat{\beta}_1 = 1(|t_{\hat{\beta}_2}| > c_n)\hat{\beta}_1 + 1(|t_{\hat{\beta}_2}| \leq c_n)\hat{\beta}_1$. Derive the asymptotic properties of $\hat{\beta}_1$ when (i) $\beta_{2,n} = \delta$ with $|\delta| > 0$ and when (ii) $\beta_{2,n} = 0$. Show that $t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1^*}{s_{\hat{\beta}_1}} \xrightarrow{d} N(0,1)$ when the null hypothesis $H_0 : \beta_1 = \beta_1^*$ is true when (i) $\beta_{2,n} = \delta$ with $|\delta| > 0$ and when (ii) $\beta_{2,n} = 0$. Conclude that $\hat{\beta}_1$ is as efficient as the “oracle estimator” that knows whether $\beta_{2,n} = 0$ despite having to learn $\beta_{2,n}$ from the data.

d. Now consider a sequence of models where $\beta_{2,n} = \frac{b_n}{\sqrt{n}}$ for some $b$ with $|b| > 0$ and $a_n > 0$ with $a_n \to \infty$ and $\frac{a_n}{\log(n)} \to 0$. Show that $\hat{\beta}_1$ is consistent. Show that $|\sqrt{n}(\hat{\beta}_1 - \beta_1^*)| \to \infty$ where $\beta_1^*$ is the true value of $\beta_1$ when $E[X_1X_2] \neq 0$ (when $\pi_{X_1} \neq 0$). Explain in words what this sequence of models captures and why this is an appropriate thought experiment for understanding the finite-sample properties of the estimator $\hat{\beta}_1$. What do the results thus far suggest about the desirability of using $\hat{\beta}_1$ in finite samples?

e. Note that the moment condition underlying the definition of $\hat{\beta}_1$ is $E[(Y - X_1\beta_1 - X_2\beta_2)X_1] = 0$. Show that this moment condition is satisfied at the true values of $\beta_1$ and $\beta_2$. Show that this moment condition does not have the orthogonality property discussed in Section 7 of the notes in that the derivative with respect to $\beta_2$ evaluated at the true parameter values is not 0.

f. Let $\tilde{\pi}_Y$ denote the least squares coefficient obtained from regressing $Y$ on $X_2$ with associated standard error estimator $s_{\tilde{\pi}_Y}$ and t-statistic for testing $\pi_Y = 0$ of $t_Y = \frac{\tilde{\pi}_Y}{s_{\tilde{\pi}_Y}}$. Similarly, let $\tilde{\pi}_{X_1}$ denote the least squares coefficient obtained from regressing $X_1$ on $X_2$ with associated standard error estimator $s_{\tilde{\pi}_{X_1}}$ and t-statistic for testing $\pi_{X_1} = 0$ of $t_{X_1} = \frac{\tilde{\pi}_{X_1}}{s_{\tilde{\pi}_{X_1}}}$. Define a new estimator $\hat{\beta}_1 = 1(|t_Y| > c_n) + 1(|t_{X_1}| > c_n)\hat{\beta}_1 + 1(|t_Y| \leq c_n)\hat{\beta}_1 + 1(|t_{X_1}| \leq c_n)\hat{\beta}_1$. Show that $\sqrt{n}(\hat{\beta}_1 - \beta_1^*) \xrightarrow{d} N(0,V)$ when $\beta_{2,n} = 0$, $\beta_{2,n} = \delta$, or $\beta_{2,n} = \frac{a_n b}{\sqrt{n}}$ regardless of the value of $E[X_1X_2]$ (and the sequence $a_n$). Suggest a consistent estimator of $V$.

g. Note that the moment condition underlying the definition of $\hat{\beta}_1$ is $E[(Y - E[Y|X_2]) - (X_1 - E[Y|X_2])\beta_1)(X_1 - E[Y|X_2])] = 0$ where $E[Y|X_2] = \pi_Y X_2$ and $E[X_1|X_2] = \pi_{X_1} X_2$. Show
that this moment condition is satisfied at the true values of $\beta_1$, $\pi_Y$, and $\pi_{X_1}$. Show that this moment condition has the orthogonality property discussed in Section 7 of the notes in that the derivative with respect to the nuisance parameters $\pi_Y$ and $\pi_{X_1}$ evaluated at the true parameter values is 0.

h. Design a simulation experiment to illustrate the potential consequences of the lack of uniformity of the estimator $\tilde{\beta}_1$ on inference. Specifically, you should be able to design a simulation experiment where the distribution of $\tilde{\beta}_1$ is strongly bimodal and size of tests based on $t_{\tilde{\beta}_1}$ is far from the nominal level. Show the robustness of $\bar{\beta}_1$ within this design in that the normal approximation provides a sensible approximation to the distribution of $\bar{\beta}_1$ across simulation replications and tests based on the $t$-statistic formed using $\tilde{\beta}_1$ and the suggested estimator of $V$ from part (f) have approximately correct size. [A test is not uniform with respect to a class of models if there are sequences of models within this class that lead to the test being size distorted even in large samples. Uniformity of inference with respect to sensible sequences of models is very important in practice as well-designed sequences are much better able to capture actual finite-sample performance of estimators.]

4. Recall that a doubly robust estimator of an average treatment effect is given by

$$\hat{\text{ATE}}_{\text{robust}} = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{D_i(Y_i - \hat{g}_1(X_i))}{\hat{e}(X_i)} - \frac{(1 - D_i)(Y_i - \hat{g}_0(X_i))}{1 - \hat{e}(X_i)} + \hat{g}_1(X_i) - \hat{g}_0(X_i) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \hat{\psi}_i.$$ 

Belloni, Chernozhukov, and Hansen (2014) show that $\sqrt{n}(\hat{\text{ATE}}_{\text{robust}} - \text{ATE}) \overset{d}{\to} N(0, V)$ where $\hat{V} = \frac{1}{n} \sum_{i=1}^{n} (\hat{\psi}_i - \hat{\text{ATE}}_{\text{robust}})^2 \overset{p}{\to} V$ when $\ell_1$-penalized estimation is used to form $\hat{g}_1(\cdot)$, $\hat{g}_0(\cdot)$, and $\hat{e}(\cdot)$ under regularity conditions including having iid data and the assumption that these functions are approximately sparse. By approximately sparse, we mean that $g_1(X_i) = X'_i\beta_1 + r_1$, $g_0(X_i) = X'_i\beta_0 + r_0$, and $e(X_i) = \Lambda(X'_i\gamma) + r_e$ with $\max\{\|\beta_1\|_0, \|\beta_0\|_0, \|\gamma\|_0\} \leq s$ where $r_1$, $r_0$, and $r_e$ are approximation errors that satisfy $\max\{E[r_1^2], E[r_0^2], E[r_e^2]\} = O(s/n)$ and $\frac{s^2 \log(p)^3}{n} \to 0$.

Consider the data in restatw.dat which contains the data used in the 401(k) example in the first two lectures. Use $e401$ (eligibility for a 401(k) plan) as the treatment variable and $\text{net.tfa}$ (net total financial assets as the dependent variable). The argument for exogeneity of a 401(k) plan
relies on conditioning on characteristics that might be associated to a person’s decision to take a job and saving preferences. Potential control variables are *age*, *inc* (income), *fsize* (family size), *educ* (years of education), *marr* (marital status), *male*, *twoearn* (part of a two-earner household), *db* (has a defined benefit pension), and *pira* (has an IRA).

a. Construct an approximating dictionary to use in $\ell_1$-penalized estimation using the control variables above. Carefully explain your choices in choosing what functions to put in the dictionary. Explain intuitively what the sparsity assumption requires in terms of this example and within the context of the dictionary you have chosen. Does it seem plausible that this assumption would be satisfied?

b. Estimate $g_1(\cdot)$ and $g_0(\cdot)$ using lasso with penalty weights that are appropriate under heteroscedasticity and penalty parameter $\lambda = 2.2\sqrt{n}\Phi^{-1}(1 - (.1/\log(n))/2p)$ where $p$ is the number of elements in your dictionary and $n$ is the appropriate sample size. (Note that $n$ will not be the same for estimating $g_0$ and $g_1$. Note that the $\lambda$ given is appropriate for solving

$$\beta_P \in \arg\min_b \frac{1}{n} \sum_{i=1}^n (y_i - x_i'b)^2 + \frac{\lambda}{n} \sum_{j=1}^p \hat{\phi}_j|b_j|.$$ 

Some lasso implementations use different scalings such as

$$\beta_P \in \arg\min_b \frac{1}{2n} \sum_{i=1}^n (y_i - x_i'b)^2 + \frac{\lambda^*}{\sum_j \hat{\phi}_j} \sum_{j=1}^p \hat{\phi}_j|b_j|$$

which would require alteration of the penalty parameter so that you are solving the same problem.) Which variables are selected to approximate each function? Do these variables make sense? Explain. Should we conclude that the selected variables are the true variables in the sense that we have captured the correct model for $E[Y_1|X]$ and $E[Y_0|X]$? Explain.

c. Estimate $e(\cdot)$ using $\ell_1$-penalized logistic regression with $\lambda = 1.1\sqrt{n}\Phi^{-1}(1 - (.1/\log(n))/2p)$ where $p$ is the number of elements in your dictionary and $n$ is the appropriate sample size. (Be careful about scaling again. This $\lambda$ is appropriate for solving

$$\gamma \in \arg\min_g \frac{1}{n} \sum_{i=1}^n \log\text{-likelihood}(y_i, x_i, g) + \frac{\lambda}{n} \sum_{j=1}^p |g_j|.$$
If the $\ell_1$-penalized logistic regression function you are using uses a different scaling, you will need to adjust $\lambda$ appropriately.) Which variables are selected? Do these variables make sense? Explain. Should we conclude that the selected variables are the true variables in the sense that we have captured the correct model for $E[D|X]$? Explain.

d. Take the selected variables for estimating $g_1$ and estimate $\widehat{g}_1(X)$ by unpenalized least squares regression of $Y$ on these selected variables in the subsample of observations with $D = 1$. Take the selected variables for estimating $g_0$ and estimate $\widehat{g}_0(X)$ by unpenalized least squares regression of $Y$ on these selected variables in the subsample of observations with $D = 0$. Take the selected variables for estimating $e(X)$ and estimate $\widehat{e}(X)$. Form fitted values for $g_1$, $g_0$, and $e$ for each observation in the data. Using these fitted values obtain $\widehat{ATE}_{robust}$ and estimate its standard error.

e. Estimate $g_1(\cdot)$ and $g_0(\cdot)$ using lasso with penalty weights that are appropriate under homoscedasticity and with penalty parameter chosen by cross-validation. Which variables are selected to approximate each function? Do these variables make sense? Explain. Do these results differ appreciably from those in part b.? Should we conclude that the selected variables are the true variables in the sense that we have captured the correct model for $E[Y_1|X]$ and $E[Y_0|X]$? Explain.

f. Estimate $e(\cdot)$ using $\ell_1$-penalized logistic regression with $\lambda$ chosen by cross-validation. Which variables are selected? Do these variables make sense? Explain. Do these results differ appreciably from those in part c.? Should we conclude that the selected variables are the true variables in the sense that we have captured the correct model for $E[D|X]$? Explain.

g. Take the estimated models from e. and f. for $g_0$, $g_1$ and $e$ (i.e. just use the coefficient estimates that come directly out of the estimation) to form fitted values for $g_1$, $g_0$, and $e$ for each observation in the data. Using these fitted values obtain $\widehat{ATE}_{robust}$ and estimate its standard error. Are these results appreciably different from those obtained in part d.? Explain the significance of the similarity or difference.