Problem Set 3

Group members: Xuchen Liu, Binqian Chai, He Jiang, Ella Lin

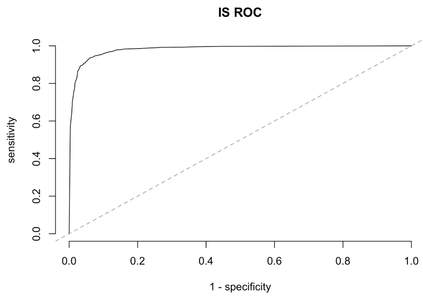
1.) We'll plot some ROC curves using the spam example from lec03–regression.nb.html. You can use the R function for plotting ROC curves, uploaded to Blackboard.

- a.) Estimate the logit specification from the lecture notes on the full data set (you can just copy the code from the lecture). Plot an ROC curve for classifying spam vs not-spam from this logit regression.

```
email <- read.csv("./data/spam.csv")
logit1 <- glm(spam ~ ., data=email, family='binomial')
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
pred1 <- predict(logit1, type="response")
source("./data/roc.R")
roc(p=pred1, y=email$spam, bty="n", main="IS ROC")
```

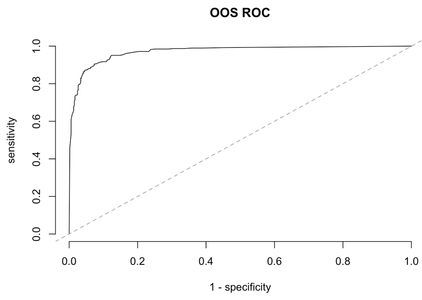


- b.) Plot an out-of-sample ROC curve by estimating the same logit specification on 4/5 of the data (sampled randomly), and evaluating classification error on the remaining 1/5 of the data.

```
set.seed(0)
test <- sample.int(length(email$spam),round(length(email$spam)/5))
logit2 <- glm(spam ~ ., data=email[-test,], family='binomial')
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
pred_oos <- predict(logit2, newdata=email[test,], type="response")
Y_oos <- email$spam[test]
roc(p=pred_oos, y=Y_oos, bty="n", main="OOS ROC")
```



2.) The data set dw_data.csv is taken from Dehejia and Wahba. It contains data on individuals who participated in a job training program (for whom treated equals TRUE) and "control" individuals from the PSID (for whom treated equals FALSE). Additional variables are: age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if maried, 0 otherwise), RE74 and RE75 (earnings in 1974 and 1975), UE74 and UE75 (indicator for unemployment in 1974 and 1975), and RE78 (earnings in 1978). The training program was done before 1978, so the last variable is post-training earnings.

- a.) Estimate a linear model with constant treatment effects by regressing RE78 on the treatment indicator and the remaining variables. Report the estimate of the average treatment effect from this regression (using the assumptions of selection-on-observables and correct specification of this regression).

```
dw <- read.csv("./data/dw_data.csv")
rel <- lm(re78 ~ ., data=dw)
coef(rel)['treatedTRUE']
```

```
## treatedTRUE
##    0.115381
```

- b.) Add interactions of the treatment indicator with the remaining variables to the regression in (a). Compute an estimate of the average treatment effect from this regression (using the assumptions of selection-on-observables and correct specification of this regression).

```
re2 <- lm(re78 ~ treated*., data=dw)
treated <- data.frame(treated = rep(TRUE, length(dw$re78)), age = dw$age, education = dw$education, black = dw$bl
ack, hispanic = dw$hispanic, married = dw$married, re74 = dw$re74, re75 = dw$re75, ue74 = dw$ue74, ue75 = dw$ue7
5, re78 = dw$re78)
control <- data.frame(treated = rep(FALSE, length(dw$re78)), age = dw$age, education = dw$education, black = dw$b
lack, hispanic = dw$hispanic, married = dw$married, re74 = dw$re74, re75 = dw$re75, ue74 = dw$ue74, ue75 = dw$ue7
5, re78 = dw$re78)
mean(predict(re2, newdata = treated) - predict(re2, newdata = control))
```

```
## [1] -8.819968
```

- c.) The data set dw_experimental_data.csv includes both the treated observations in the dw_data.csv data set and a set of true control individuals, such that treatment was randomized among this population. Parts (a) and (b) can thus be considered an attempt to recover the outcome of this experiment using observational data where a true control group is not available. Using the data set dw_experimental_data.csv, estimate the average treatment effect by taking the difference in means between treated and control groups. Is the answer similar to (a) and (b)?

```
dw_experimental <- read.csv("./data/dw_experimental_data.csv")
treated_group <- mean(dw_experimental[dw_experimental$treated== TRUE, 're78'])
control_group <- mean(dw_experimental[dw_experimental$treated== FALSE, 're78'])
treated_group-control_group
```

```
## [1] 1.794342
```

The answer is not similar to (a) and (b).

3.) We'll run a placebo test for the RD estimator using the Lee data.

- a.) Run a local linear regression estimator (with uniform kernel) as described in the lecture notes. Use h = 5 as the bandwidth.

```
lee08 <- read.csv("./data/lee08.csv")
bw <- 5
local_linear_model <- lm(voteshare ~ I(margin>=0)*margin, data = lee08, subset = (abs(margin) <= bw))
summary(local_linear_model)
```

```
##
## Call:
## lm(formula = voteshare ~ I(margin >= 0) * margin, data = lee08,
##     subset = (abs(margin) <= bw))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.588  -5.505  -0.760   3.560  47.743
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               46.9239     1.3869  33.833   <2e-16 ***
## I(margin >= 0)TRUE         4.8613     1.8897   2.573   0.0103 *
## margin                     0.9029     0.4795   1.883   0.0602 .
## I(margin >= 0)TRUE:margin  0.0153     0.6439   0.024   0.9811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.14 on 606 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1612
## F-statistic: 40.01 on 3 and 606 DF,  p-value: < 2.2e-16
```

b.) Form placebo estimates β_0^+ at cutoffs c_1 ,…,c_B given by −80,−79,…,−7,−6,−5 and 5,6,7,…,79,80. Plot a histogram of the placebo estimates β_0^+ along with a vertical line at the estimate computed in part (a).

```
c <- c(-80:-5, 5:80)
estimates <- c(rep(0, length(c)))
for (i in (1:length(c))) {
  model <- lm(voteshare ~ I(margin>=c[i])*margin, data = lee08, subset = (abs(margin-c[i]) <= bw))
  estimates[i] <- coef(model)['I(margin >= c[i])TRUE']
}
hist(estimates, main = "Histogram of Placebo Estimates", xlab = "Placebo Estimates", ylab = "Frequency", breaks =
30)
abline(v = coef(local_linear_model)['I(margin >= 0)TRUE'], col = 'red', lwd = 1)
```