Biostats - Causal Inference

Rob Leonard (robleonard@tamu.edu (mailto:robleonard@tamu.edu))

1) Real Earnings

There are 2 sets of assumptions: First, the assumptions for valid linear regression: The response and the explanatory variables are linearly related.
Residuals are iid Normal (0,σ^2) where the variance is constant.

Second, the assumptions for **causality:**
Consistency: the counter factual outcomes are partially known, and we have response data for both the treatment and non treatment group.
Exchangeability: the counter factual outcomes and assignment of exposure are independent conditional on the confounder.
Positivity: Both pr(X=1|Z=z) and pr(X=0|Z=z) are positive for every value of z.

```
library(qte)
data(lalonde)
mydata1           = lalonde.exp.panel
mydata1$y        = log(1+mydata1$re)   # set up log of real earnings - but we have some with 0 w
ages so use 1+re.
lm1               = lm(re~treat+age+education+black+hispanic+married+nodegree+treat*age+treat*ed
ucation+
                       treat*black+treat*hispanic+treat:married+treat:nodegree, data=mydata1)
summary(lm1)
```

```
##
## Call:
## lm(formula = re ~ treat + age + education + black + hispanic +
##     married + nodegree + treat * age + treat * education + treat *
##     black + treat * hispanic + treat:married + treat:nodegree,
##     data = mydata1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6568  -2706  -2226   1099  57466
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3665.679   2182.062   1.680  0.09321 .
## treat           -3034.494   3245.481  -0.935  0.34996
## age                 6.469     28.584   0.226  0.82099
## education          32.916    153.294   0.215  0.83001
## black           -2202.979    806.269  -2.732  0.00637 **
## hispanic        -1600.777    986.841  -1.622  0.10502
## married          1200.986    558.240   2.151  0.03163 *
## nodegree          350.340    659.533   0.531  0.59537
## treat:age           7.369     44.527   0.165  0.86858
## treat:education   190.931    220.650   0.865  0.38702
## treat:black      2327.340   1130.961   2.058  0.03980 *
## treat:hispanic   2613.803   1573.819   1.661  0.09699 .
## treat:married     896.791    830.124   1.080  0.28020
## treat:nodegree  -1113.969    966.803  -1.152  0.24944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5459 on 1321 degrees of freedom
## Multiple R-squared:  0.03007,    Adjusted R-squared:  0.02053
## F-statistic: 3.151 on 13 and 1321 DF,  p-value: 0.0001147
```

```
# calculate average treatment effect
mydata1x1         = mydata1
mydata1x1$treat   = 1
yhat1             = predict(lm1,newdata=mydata1x1)
mydata1x0         = mydata1
mydata1x0$treat   = 0
yhat0             = predict(lm1,newdata=mydata1x0)
ATE1              = yhat1 - yhat0
mean(ATE1)
```

```
## [1] 548.493
```

```
# get std error of the treatment effect and a 95% CI using the bootstrap method
B                 = 10000
n                 = 1335
ATEboot           = rep(0,1335)

for (b in 1:B){
  bootdata1       = mydata1x1[sample(nrow(mydata1x1), n, replace=TRUE),]
  bootdata0       = mydata1x0[sample(nrow(mydata1x0), n, replace=TRUE),]
  treat1          = predict(lm1, newdata=bootdata1)
  treat0          = predict(lm1, newdata=bootdata0)
  ATEboot[b]      = mean(treat1-treat0)
}
ATEboot           = sort(ATEboot)
ATEboot[250]
```

```
## [1] 476.0364
```

```
ATEboot[9750]
```

```
## [1] 619.6703
```

The estimated average treatment effect (increase in real earnings after participating in the training program vs. not participating) is **$548.49**, and the 95% CI is **(476.0363839, 619.6702975).**

2) Chemotherapy and Survival

IPTW Method First

```
# setup the data
library(stdReg)
library(AF)
```

```
## Loading required package: survival
```

```
## Loading required package: drgee
```

```
## Loading required package: data.table
```

```
## Loading required package: ivtools
```

```
library(survival)
data(rott2)
mydata2          = rott2
mydata2$chemoind = as.numeric(rott2$chemo=="yes")  # set up yes as 1 for treatment
mydata2$lpr      = log(1+mydata2$pr)
mydata2$ler      = log(1+mydata2$er)
# calculate the weights using IPTW
out2             = glm(chemoind~factor(meno)+factor(size)+factor(grade)+nodes+lpr+ler, family =
 binomial, data=mydata2)
summary(out2)
```

```
##
## Call:
## glm(formula = chemoind ~ factor(meno) + factor(size) + factor(grade) +
##     nodes + lpr + ler, family = binomial, data = mydata2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2281  -0.8028  -0.2685  -0.2235   2.7215
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -3.76344    0.21933 -17.159   <2e-16 ***
## factor(meno)pre        2.69262    0.14558  18.496   <2e-16 ***
## factor(size)>20-50mmm  0.10381    0.11541   0.899   0.3684
## factor(size)>50mm      0.28048    0.19436   1.443   0.1490
## factor(grade)3         0.06033    0.12111   0.498   0.6184
## nodes                  0.13494    0.01265  10.665   <2e-16 ***
## lpr                    0.05443    0.03223   1.689   0.0912 .
## ler                   -0.01886    0.03738  -0.505   0.6138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2938.4  on 2981  degrees of freedom
## Residual deviance: 2272.1  on 2974  degrees of freedom
## AIC: 2288.1
##
## Number of Fisher Scoring iterations: 5
```

```
esttreatprob     = 1/(1+exp(-out2$linear.predictors))
our.wt           = mydata2$rfi/esttreatprob + (1-mydata2$rfi)/(1-esttreatprob)
summary(our.wt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.020   1.060   1.696   7.903   7.620  49.182
```

```
sum(our.wt)  # so a max 49 weighting is well less than 1%, weights seem valid.
```

```
## [1] 23567.76
```

```
# calculate the survival - first for the treatment group
km.IPTW.tr       = survfit(with(mydata2, Surv(rf, rfi))~1, data=mydata2, subset=(chemoind==1), w
eights=our.wt)
print(km.IPTW.tr, print.rmean=TRUE)
```

```
## Call: survfit(formula = with(mydata2, Surv(rf, rfi)) ~ 1, data = mydata2,
##     weights = our.wt, subset = (chemoind == 1))
##
##    records           n      events     *rmean *se(rmean)     median    0.95LCL
##     580.00    2100.24    1709.58      66.09       1.33      46.69      37.72
##    0.95UCL
##      53.85
##     * restricted mean with upper limit =  206
```

```
# control group
km.IPTW.con      = survfit(with(mydata2, Surv(rf, rfi))~1, data=mydata2, subset=(chemoind==0), w
eights=our.wt)
print(km.IPTW.con, print.rmean=TRUE)
```

```
## Call: survfit(formula = with(mydata2, Surv(rf, rfi)) ~ 1, data = mydata2,
##     weights = our.wt, subset = (chemoind == 0))
##
##    records           n      events     *rmean *se(rmean)     median    0.95LCL
##   2.40e+03   2.15e+04   2.00e+04   4.90e+01   2.94e-01   3.60e+01   3.33e+01
##    0.95UCL
##   3.88e+01
##     * restricted mean with upper limit =  231
```

```
# add bootstrap to get 95% CI
n                = nrow(mydata2)
B                = 200
boot.mean.tr     = boot.median.tr = boot.mean.con = boot.median.con = rep(0,200)
for (b in 1:B){
  bootdata           = mydata2[sample(nrow(mydata2), n, replace=TRUE),]
  maxtime            = max(bootdata$rf)
  # get propensity scores and weights
  bootout2           = glm(chemoind~factor(meno)+factor(size)+factor(grade)+nodes+lpr+ler, famil
y = binomial,                                          data=bootdata)
  bootesttreatprob   = 1/(1+exp(-bootout2$linear.predictors))
  boot.wt            = bootdata$rfi/bootesttreatprob + (1-bootdata$rfi)/(1-bootesttreatprob)
  # estimate mean and median survival times
  boot.km.tr         = survfit(with(bootdata, Surv(rf, rfi))~1, data=bootdata, subset=(chemoind=
=1),                                                 weights=boot.wt)
  boot.mean.tr[b]    = survival:::survmean(boot.km.tr, rmean=maxtime)[[1]]["*rmean"]
  boot.median.tr[b]  = quantile(boot.km.tr, prob = 0.5)[1]
  boot.km.con        = survfit(with(bootdata, Surv(rf, rfi))~1, data=bootdata, subset=(chemoind=
=0),                                                 weights=boot.wt)
  boot.mean.con[b]   = survival:::survmean(boot.km.con, rmean=maxtime)[[1]]["*rmean"]
  boot.median.con[b] = quantile(boot.km.con, prob = 0.5)[1]
}
boot.median.tr  = unlist(boot.median.tr)  # unlist the median values
boot.median.con = unlist(boot.median.con)

trteff.mean     = boot.mean.tr-boot.mean.con
trteff.median   = boot.median.tr-boot.median.con
trteff.mean     = sort(trteff.mean)
trteff.mean[5]
```

```
## [1] 12.11336
```

```
trteff.mean[195]
```

```
## [1] 26.4987
```

```
trteff.median   = sort(trteff.median)
trteff.median[5]
```

```
##       50
## 0.295689
```

```
trteff.median[195]
```

```
##      50
## 18.2998
```

Using the IPTW method, the **mean** survival time for **chemo** patients was **66 months** and for **non chemo** patients was **49 months.** The average treatment effect on the mean survival time for taking chemo is 17 months with a **95% confidence interval of (12.1133568, 26.4987036).**

The **median** survival time for **chemo** patients was **46.7 months** and for **non chemo** patients was **36 months.** The average treatment effect on the median survival time for taking chemo was 10.7 months with a **95% CI of (0.295689, 18.299797)**.

Now use CBPS Method

```
library(CBPS)
```

```
## Loading required package: MASS
```

```
## Loading required package: MatchIt
```

```
## Loading required package: nnet
```

```
## Loading required package: numDeriv
```

```
## Loading required package: glmnet
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.0-2
```

```
## CBPS: Covariate Balancing Propensity Score
## Version: 0.21
## Authors: Christian Fong [aut, cre],
##   Marc Ratkovic [aut],
##   Kosuke Imai [aut],
##   Chad Hazlett [ctb],
##   Xiaolin Yang [ctb],
##   Sida Peng [ctb]
```

```
out2CBPS       = CBPS(chemoind~factor(meno)+factor(size)+factor(grade)+nodes+lpr+ler, data=mydat
a2)
```

```
## [1] "Finding ATT with T=1 as the treatment.  Set ATT=2 to find ATT with T=0 as the treatment"
```

```
balance(out2CBPS)
```

```
## $balanced
##                           0.mean    1.mean 0.std.mean 1.std.mean
## factor(meno)pre       0.8436616 0.8465517  1.6993308  1.7051522
## factor(size)>20-50mmm 0.4533252 0.4534483  0.9147654  0.9150138
## factor(size)>50mm     0.1344049 0.1344828  0.4441276  0.4443849
## factor(grade)3        0.7346658 0.7344828  1.6618457  1.6614317
## nodes                 4.1924655 4.1982759  0.9563445  0.9576699
## lpr                   3.6875373 3.6896300  1.6495798  1.6505159
## ler                   3.3818169 3.3789051  1.6635266  1.6620943
##
## $original
##                           0.mean    1.mean 0.std.mean 1.std.mean
## factor(meno)pre       0.34179850 0.8465517  0.6884617  1.7051522
## factor(size)>20-50mmm 0.42797669 0.4534483  0.8636146  0.9150138
## factor(size)>50mm     0.09408826 0.1344828  0.3109053  0.4443849
## factor(grade)3        0.73355537 0.7344828  1.6593339  1.6614317
## nodes                 2.35345545 4.1982759  0.5368474  0.9576699
## lpr                   3.36257284 3.6896300  1.5042104  1.6505159
## ler                   3.85607031 3.3789051  1.8968134  1.6620943
```

```
# there are some unbalanced confounders, so rerun with new weightings
km.CBPS.tr     = survfit(with(mydata2, Surv(rf, rfi))~1, data=mydata2, subset=(chemoind==1),
                     weights=out2CBPS$weights)
print(km.CBPS.tr, print.rmean=TRUE)
```

```
## Call: survfit(formula = with(mydata2, Surv(rf, rfi)) ~ 1, data = mydata2,
##     weights = out2CBPS$weights, subset = (chemoind == 1))
##
##    records          n     events     *rmean *se(rmean)     median    0.95LCL
##    580.000      1.000      0.581    101.211     88.058     73.692     60.057
##    0.95UCL
##     90.743
##     * restricted mean with upper limit =  206
```

```
km.CBPS.con    = survfit(with(mydata2, Surv(rf, rfi))~1, data=mydata2, subset=(chemoind==0),
                     weights=out2CBPS$weights)
print(km.CBPS.con, print.rmean=TRUE)
```

```
## Call: survfit(formula = with(mydata2, Surv(rf, rfi)) ~ 1, data = mydata2,
##     weights = out2CBPS$weights, subset = (chemoind == 0))
##
##    records          n     events     *rmean *se(rmean)     median    0.95LCL
##   2402.000      1.000      0.577    104.962    114.659     68.008     53.224
##    0.95UCL
##     89.593
##     * restricted mean with upper limit =  231
```

```
# Run bootstrap to get 95% CI's
n                     = nrow(mydata2)
B                     = 200
boot.mean.trCB        = boot.median.trCB = boot.mean.conCB = boot.median.conCB = rep(0,200)
for (b in 1:B){
  bootdataCB          = mydata2[sample(nrow(mydata2), n, replace=TRUE),]
  maxtime             = max(bootdata$rf)
  # get weights
  bootout2CB          = CBPS(chemoind~factor(meno)+factor(size)+factor(grade)+nodes+lpr+ler, da
ta=bootdataCB)
  boot.wtCB           = bootout2CB$weights
  # estimate mean and median survival times
  boot.km.trCB        = survfit(with(bootdataCB, Surv(rf, rfi))~1, data=bootdataCB, subset=(che
moind==1),                                        weights=boot.wtCB)
  boot.mean.trCB[b]   = survival:::survmean(boot.km.trCB, rmean=maxtime)[[1]]["*rmean"]
  boot.median.trCB[b] = quantile(boot.km.trCB, prob = 0.5)[1]
  boot.km.conCB       = survfit(with(bootdataCB, Surv(rf, rfi))~1, data=bootdataCB, subset=(che
moind==0),                                        weights=boot.wtCB)
  boot.mean.conCB[b]  = survival:::survmean(boot.km.conCB, rmean=maxtime)[[1]]["*rmean"]
  boot.median.conCB[b] = quantile(boot.km.conCB, prob = 0.5)[1]
}
```

```
boot.median.trCB       = unlist(boot.median.trCB)  # unlist the median values
boot.median.conCB      = unlist(boot.median.conCB)

trteff.meanCB          = boot.mean.trCB-boot.mean.conCB
trteff.medianCB        = boot.median.trCB-boot.median.conCB
trteff.meanCB          = sort(trteff.meanCB)
trteff.meanCB[5]


## [1] -6.177012


trteff.meanCB[195]
```

```
## [1] 15.59448
```

```
trteff.medianCB        = sort(t
trteff.medianCB[5]
```

```
##        50
## -18.62833
```

```
trteff.medianCB[195]
```

```
##      50
## 24.9692
```

Using the **CBPS** method, the **mean** survival time for **chemo** patients was **101.2 months** and for **non chemo** patients was **105 months.** The average treatment effect on the mean survival time for taking chemo is **-3.8 months** with a **95% confidence interval of (-6.1770117, 15.5944756).**

The **median** survival time for **chemo** patients was **73.7 months** and for **non chemo** patients was **68 months.** The average treatment effect on the median survival time for taking chemo was **5.7 months** with a **95% CI of (-18.628334, 24.9692)**.