Problem Set 5

Research Design for Causal Inference
Due: May 12, 2015

For this Problem Set, you'll re-analyze a subset of the data from:

> Blattman, Christopher, and J. Annan. 2010. "The Consequences of Child Soldiering," *Review of Economics & Statistics*, 92 (4): 882-898. [Available via MIT Press]

The study consists of a panel survey of male youth in several war-afflicted areas of Uganda. The authors attempt to estimate the impact of forced military service – in this case, abduction into the Lord's Resistance Army – on young men's educational and labor market outcomes. Blattman and Annan describe the abductions (2010: 883):

> Abduction was large-scale and seemingly indiscriminate; 60,000 to 80,000 youth are estimated to have been abducted and more than a quarter of males currently aged 14 to 30 in our study region were abducted for at least two weeks. Most were abducted after 1996 and from one of the Acholi districts of Gulu, Kitgum, and Pader.
>
> Youth were typically taken by roving groups of 10 to 20 rebels during night raids on rural homes. Adolescent males appear to have been the most pliable, reliable and effective forced recruits, and so were disproportionately targeted by the LRA. Youth under age 11 and over 24 tended to be avoided and had a high probability of immediate release. Lengths of abduction ranged from a day to ten years, averaging 8.9 months in our sample. Youth who failed to escape were trained as fighters and, after a few months, received a gun. Two thirds of abductees were forced to perpetrate a crime or violence. A third eventually became fighters, and a fifth were forced to murder soldiers, civilians, or even family members in order to bind them to the group, to reduce their fear of killing, and to discourage disobedience.

Assignment to "treatment" (a.k.a. abduction) was not *truly* randomized, but (according to Blattman & Annan) was "as-if" randomized, supposedly resulting in conditional independence of treatment assignment on any of the observed or unobserved covariates. The overarching goals here are to (1) assess the as-if random assumption; and (2) estimate treatment effects using multiple techniques in order to address any potential breakdowns in the as-if random assumption.

The dataset is available as a comma-separated value (.csv) file at: http://aaronshaw.org/teaching/2015/causal/data/blattman.csv. The variables included in the dataset are described in Table 1. Note that `educ`, `distress`, and `log.wage` are all post-treatment outcomes.

Table 1: Variables in Blattman & Annan (2010) dataset

Variable name	Definition
`abd`	1 if respondent was abducted by the LRA (treatment)
`c_ach -- c_pal`	Location indicators corresponding to a geographic sub-district
`age`	Age (years)
`fthr_ed`	Father's education (years)
`mthr_ed`	Mother's education (years)
`orphan96`	Indicator if parents died before 1997
`hh_fthr_frm`	Indicator if father is a farmer
`hh_size96`	Household size in 1996
`educ`	Respondent's education (years)
`distress`	Emotional distress (index 0–15)
`log.wage`	Log of average daily wage over previous 4 weeks

Question 1 – Observational Study of Child Soldiering

Part a Use OLS regression to calculate estimates of the Average Treatment Effect of abduction on education. Generate an estimate with and without covariate adjustment. Interpret the results in a few sentences.

```
# Load the dataset:
d <-
read.csv('http://aaronshaw.org/teaching/2015/causal/data/blattman.csv')

# Here's an estimate without covariance adjustment
summary(lm(educ ~ abd, data=d))

##
```

```
## Call:
## lm(formula = educ ~ abd, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4158 -1.8203 -0.4158  2.1797  8.5842
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.4158     0.1722  43.072  < 2e-16 ***
## abd          -0.5954     0.2180  -2.731  0.00647 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.876 on 739 degrees of freedom
## Multiple R-squared:  0.00999,Adjusted R-squared:  0.00865
## F-statistic: 7.457 on 1 and 739 DF,  p-value: 0.00647

# Now for a more fully specified model with covariance adjustment:
# To make things easier, I'm going to create variables that store
# the names of my covariates and other variables respectively:

not.covars <- c("abd", "educ", "distress", "log.wage")
covars <- names(d)[!names(d) %in% not.covars]

# Now, I'll create a formula that I can pass into a regression model:
f.educ <- as.formula(paste("educ ~ abd + ", paste(covars, collapse = "+") ) )

summary(lm(f.educ, data=d))

##
## Call:
## lm(formula = f.educ, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0168 -1.8836 -0.2596  1.6959  7.6601
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.25689    0.70413   8.886  < 2e-16 ***
```

```
## abd          -0.70447    0.21738  -3.241 0.001246 **
## C_ach        -1.53242    0.39918  -3.839 0.000134 ***
## C_akw        -0.46191    0.41597  -1.110 0.267180
## C_ata        -1.54609    0.40445  -3.823 0.000143 ***
## C_kma        -0.75832    0.40413  -1.876 0.060994 .
## C_oro        -1.77058    0.45812  -3.865 0.000121 ***
## C_pad        -1.19245    0.41283  -2.888 0.003986 **
## C_paj        -0.52873    0.40452  -1.307 0.191603
## C_pal              NA         NA      NA       NA
## age           0.04536    0.02063   2.199 0.028192 *
## fthr_ed       0.14088    0.03066   4.595 5.09e-06 ***
## mthr_ed       0.05561    0.03696   1.504 0.132911
## orphan96      0.22994    0.38193   0.602 0.547319
## hh_fthr_frm -0.30757    0.35666  -0.862 0.388780
## hh_size96     0.06476    0.02605   2.486 0.013140 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.743 on 726 degrees of freedom
## Multiple R-squared:  0.1151,Adjusted R-squared:  0.09807
## F-statistic: 6.747 on 14 and 726 DF,  p-value: 4.665e-13

# Based on this model (and taking the covariate-adjusted
# coefficient as likely to be a more precise point-estimate),
# abduction results in an average reduction of education of about 0.7
# years.
```

Part b This study is not the product of a "pure" natural experiment (treatment was not randomly assigned). Given that fact (and what you know about observational studies in general), what are some of the potential threats to the validity of your estimates in Part a?

> The biggest threat to the validity of the estimates reported in Part a derives from some form of omitted variable bias in Blattman & Annan's data. Despite their claims that abduction was "as-if" random, it is possible that some missing covariate explains the systematic variation between abducted and not-abducted young men. No statistical test or analysis could detect or reject this possibility.
>
> Another related threat concerns the possibility that the abducted individuals are systematically different from their peers on observed characteristics. If this is the case, statistical techniques *can* help recover unbiased estimates by trim-

ming the dataset and weighting to ensure balance on covariates across the treatment and control groups.

Part c Assess the covariate balance between the abducted and non-abducted youth respondents in the study for all pre-treatment covariates. Include any descriptive statistics you think are relevant and report the results of any statistical tests you use in your assessment.

```
# Here's where that variable storing covariate names really becomes
# helpful:

# Uncomment this lapply() command and run it. The output is long
# so I didn't include it here.
# lapply(d[,covars], summary)

# Uncomment this one too - descriptive comparisons across
# treatment and control.
# lapply(d[,covars], function(x) table(x, d$abd))

# Here are many t-tests
lapply(d[,covars], function(x){ t.test(x[d$abd == 1], x[d$abd == 0])})

## $C_ach
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = 1.5322, df = 642.254, p-value = 0.126
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01097864  0.08894727
## sample estimates:
## mean of x mean of y
## 0.1536797 0.1146953
##
##
## $C_akw
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
```

```
## t = 3.3756, df = 709.386, p-value = 0.0007767
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03311718 0.12519404
## sample estimates:
##  mean of x  mean of y
## 0.15800866 0.07885305
##
##
## $C_ata
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = -3.5303, df = 467.47, p-value = 0.0004563
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.15187343 -0.04325761
## sample estimates:
## mean of x mean of y
## 0.0995671 0.1971326
##
##
## $C_kma
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = 1.2996, df = 633.747, p-value = 0.1942
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.01698372  0.08345488
## sample estimates:
## mean of x mean of y
## 0.1515152 0.1182796
##
##
## $C_oro
##
##  Welch Two Sample t-test
##
```

```
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = -3.6596, df = 419.933, p-value = 0.0002848
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.12950628 -0.03899905
## sample estimates:
##  mean of x  mean of y
## 0.05194805 0.13620072
##
##
## $C_pad
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = -0.0263, df = 584.836, p-value = 0.9791
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04939763  0.04809428
## sample estimates:
## mean of x mean of y
## 0.1212121 0.1218638
##
##
## $C_paj
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = 1.92, df = 658.344, p-value = 0.05529
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001078951  0.096223949
## sample estimates:
## mean of x mean of y
## 0.1515152 0.1039427
##
##
## $C_pal
##
##  Welch Two Sample t-test
```

```
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = -0.6613, df = 559.023, p-value = 0.5087
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.06542329  0.03246700
## sample estimates:
## mean of x mean of y
## 0.1125541 0.1290323
##
##
## $age
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = 3.239, df = 595.876, p-value = 0.001266
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.4783922 1.9521343
## sample estimates:
## mean of x mean of y
##  21.36580  20.15054
##
##
## $fthr_ed
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = -1.1125, df = 572.987, p-value = 0.2664
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.8408032  0.2327410
## sample estimates:
## mean of x mean of y
##  5.764069  6.068100
##
##
## $mthr_ed
##
```

```
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = -1.7226, df = 516.434, p-value = 0.08556
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.85949991  0.05639979
## sample estimates:
## mean of x mean of y
##  2.093074  2.494624
##
##
## $orphan96
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = 0.1316, df = 593.215, p-value = 0.8953
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03693083  0.04223736
## sample estimates:
##  mean of x  mean of y
## 0.07792208 0.07526882
##
##
## $hh_fthr_frm
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = -0.523, df = 611.621, p-value = 0.6012
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05411539  0.03135321
## sample estimates:
## mean of x mean of y
## 0.9025974 0.9139785
##
##
## $hh_size96
```

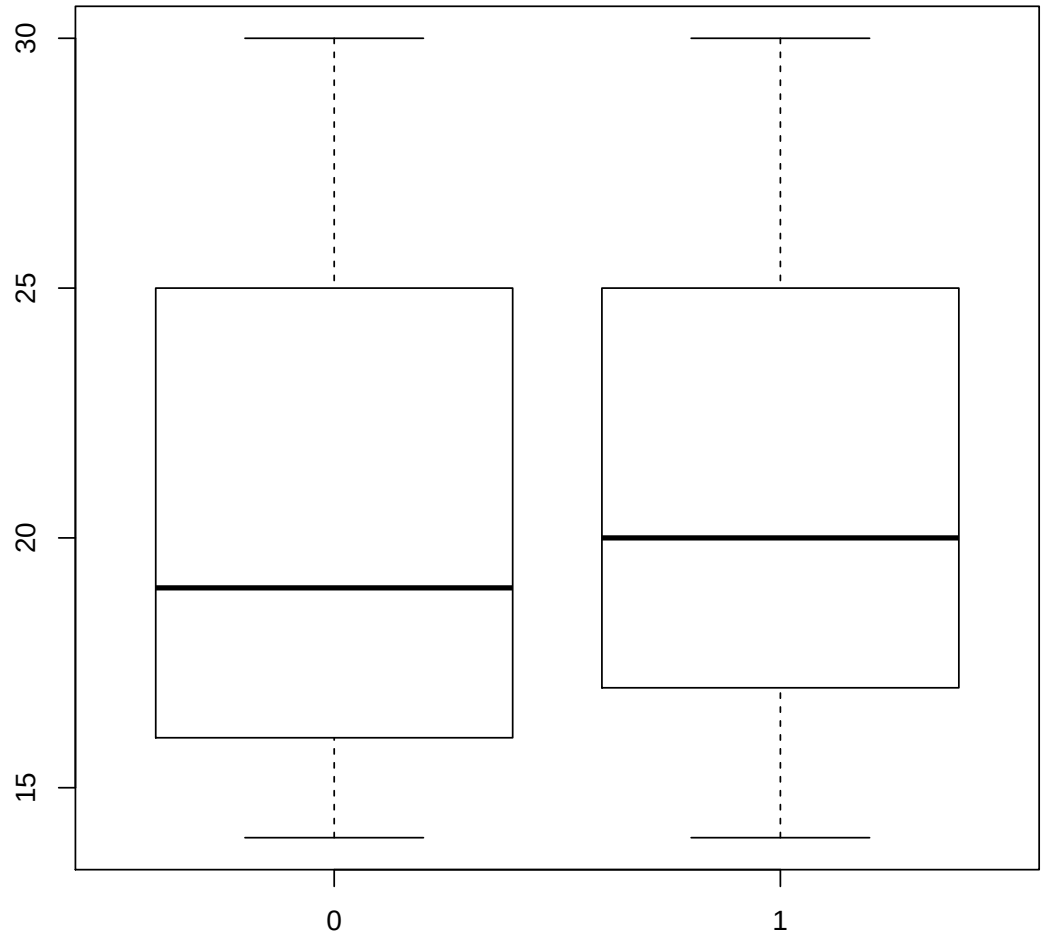
```
##
##  Welch Two Sample t-test
##
## data:  x[d$abd == 1] and x[d$abd == 0]
## t = -1.9033, df = 536.863, p-value = 0.05754
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.22911754  0.01942233
## sample estimates:
## mean of x mean of y
##  8.090493  8.695341

# The Kolmogorov-Smirnov ("ks") test is a very useful way
# to compare the entire distributions of continuous variables.
# If the D statistic is large, the null (that the distributions
# are the same) will be rejected (with a corresponding p value).

# Uncomment the lines below to run KS tests on all
# continuous variables:
# cont.covars <- c("age", "fthr_ed", "mthr_ed", "hh_size96")
# lapply(d[,cont.covars], function(x){ ks.test(x[d$abd == 1], x[d$abd == 0])})

# And Chi-square tests for binary covariates (again, commented out)
# binary.covars <- covars[!covars %in% cont.covars]
# lapply(d[,binary.covars], function(x){ chisq.test(table(x, d$abd))})
```

The results of descriptive comparisons, t-tests, Chi-square tests, and Kolmogorov-Smirnov tests indicate that several of the covariates are distributed significantly differently across the abducted and not-abducted groups.

The fact that so many of the covariates are not evenly distributed between the treatment and control conditions implies that treatment assignment cannot be considered independent of outcomes conditional on other covariates (recall: this is what you would expect to get from true randomization and provides the "equality of expectation" necessary for causal inference within the potential outcomes framework). As a result, even though treatment assignment may have been "as-if" random, naive estimates of average treatment effects that do not account for these imbalances will likely be biased (irrespective of whether selection into treatment has occurred only on observables or not).

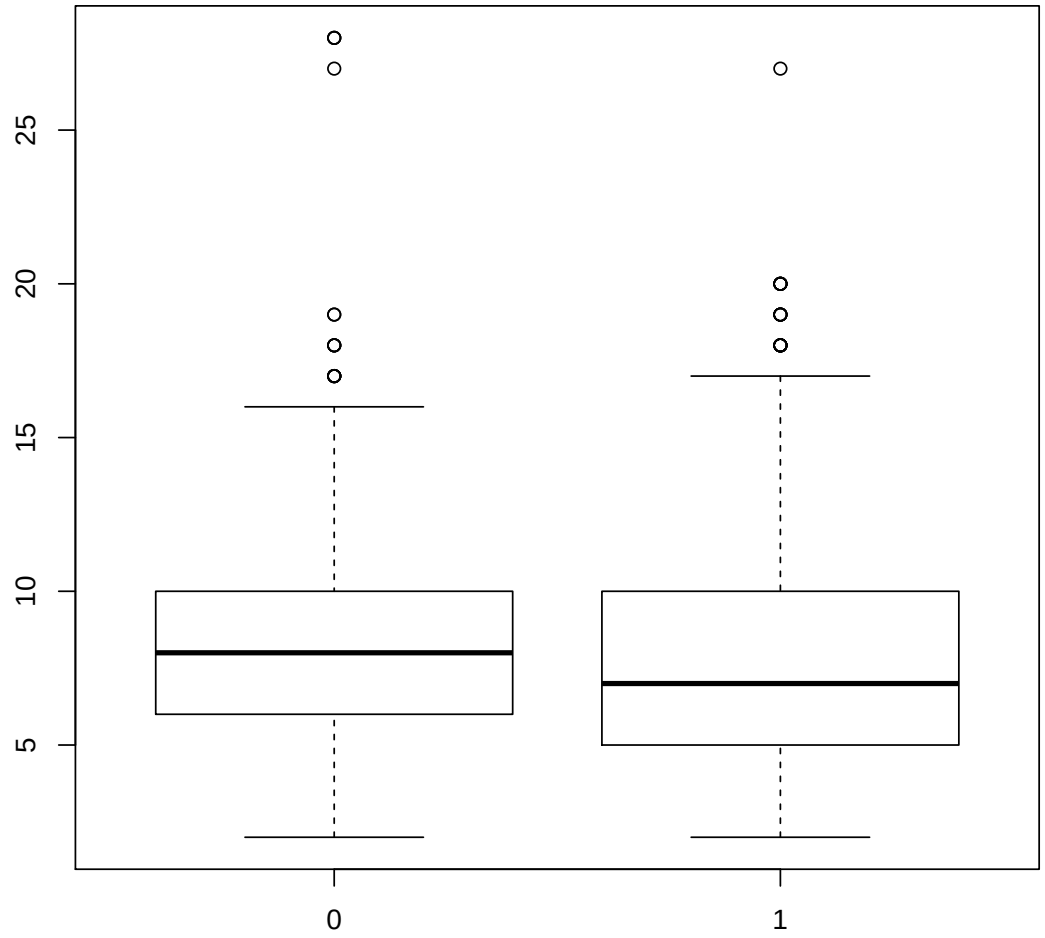
Part d Visually assess balance on a few covariates (of your choice) by using a box-plot. The default graphics command for a box-plot in R is: `boxplot(variable ~ categories)`, where `categories` is some categorical variable and `variable` is some continuous variable you are comparing across each category in `categories`.

```
# Using R's default graphics engine:
boxplot(d$age ~ d$abd)
```

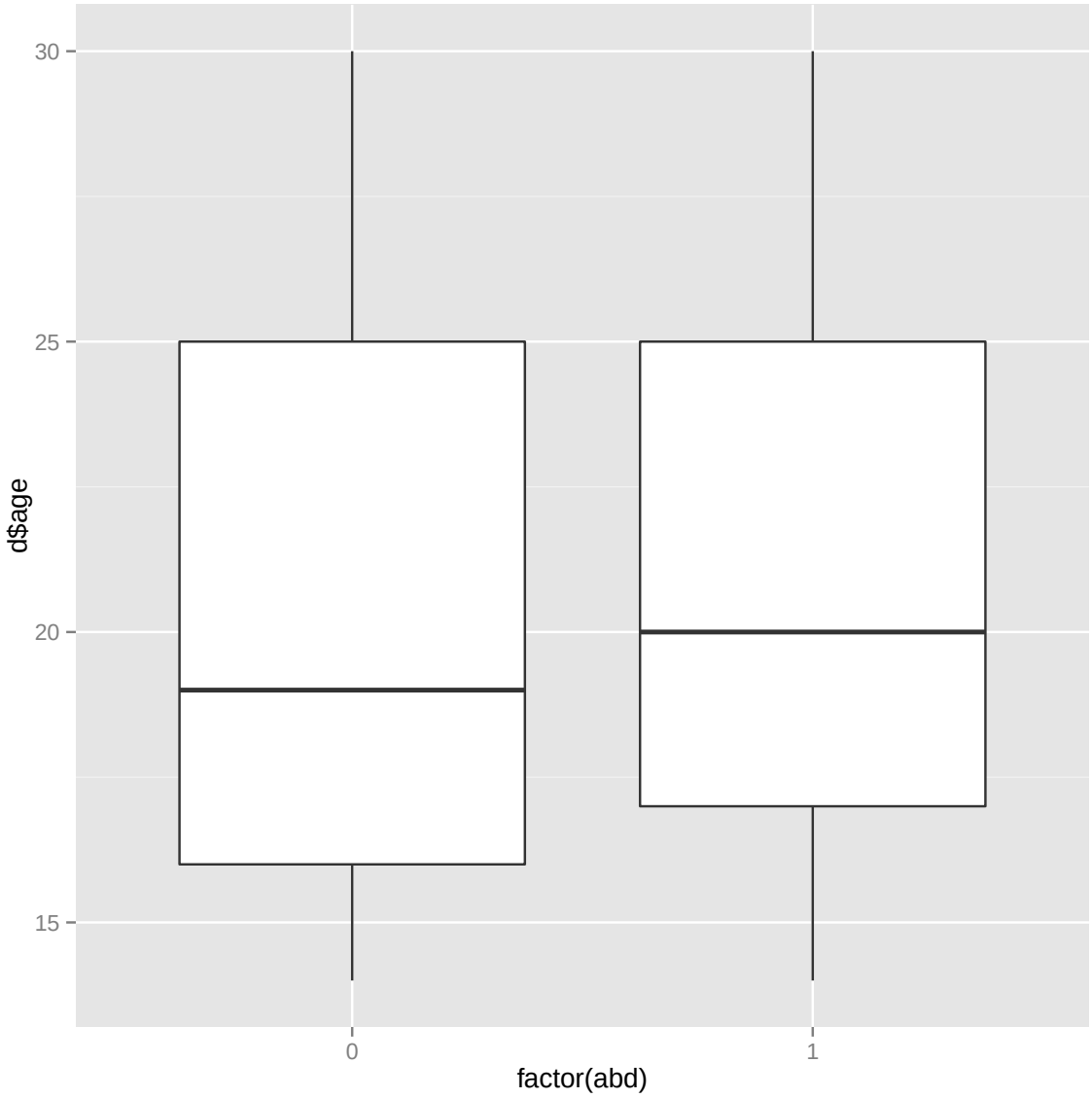



```
boxplot(d$hh_size96 ~ d$abd)
```

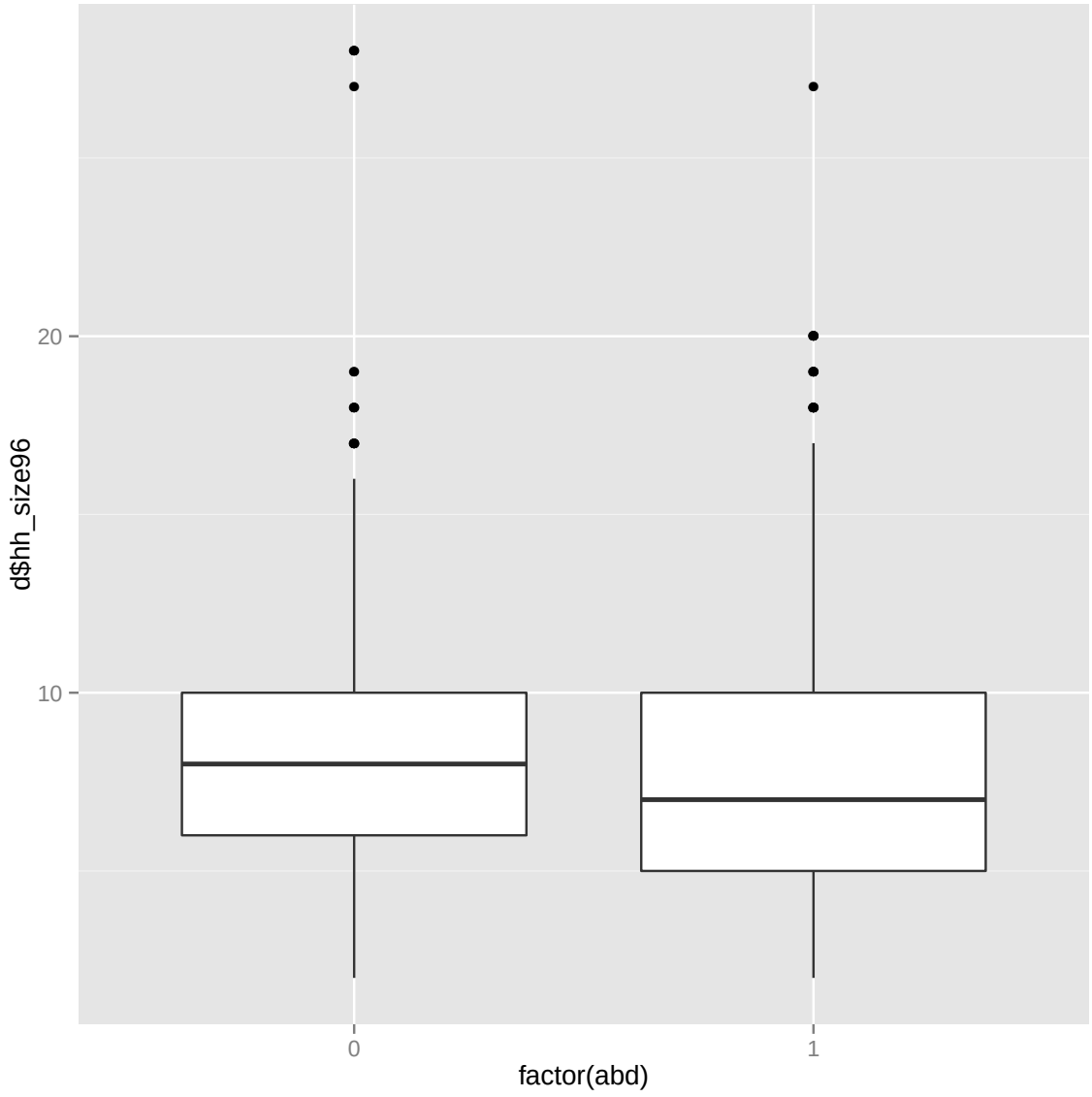
```
# Using ggplot2
library(ggplot2)
```




```
ggplot(d, aes(factor(abd), d$age)) + geom_boxplot()
```




```
ggplot(d, aes(factor(abd), d$hh_size96)) + geom_boxplot()
```



QUESTION 2 – PROPENSITY SCORE WEIGHTING

For this question, you'll use an inverse-propensity score weighting technique very similar to that employed by Blattman and Annan to attempt to improve the precision and validity of the estimates you calculated in Question 1, Part a.

The intuition guiding this effort comes from the work of Don Rubin and Paul Rosenbaum, who in a 1983 article introduced the "propensity score" as a means of generating improved balance between treated and un-treated units conditional on some combination observed covariates. *The propensity score (p-score) is an estimate of the probability that any unit in an observational study was selected into either treatment or control, given some combi-*

nation of observed covariates. For details on propensity score adjustment and the properties of different propensity score adjusted estimators, see the Hirano, Imbens, and Ritter (2003) paper cited by Blattman and Annan.

To generate improved estimates from Question 1, I ask you to (1) calculate the propensity score for every unit using a logistic regression model; (2) "trim" the dataset to remove units with extreme p-score values; (3) assess covariate balance in the trimmed dataset; (4) estimate treatment effects using a regression model that incorporates the inverse propensity scores as weights.

Part a Use a logistic regression model to estimate the probability that subjects in Blattman & Annan's study were treated (abducted). A generic R command for performing logistic regression is as follows:

```
model <- glm( y ~ x1 + x2 + ... + xN, data = DATA,
         family = binomial(link = logit))
```

where y is the dependent variable; each x is an independent variable and `DATA` is your data frame. Two comments: (1) in theory, the more covariates you use in calculating your p-score, the better; however, in practice you may discover that it makes sense to eliminate some covariates from your p-score model if they include many missing values or empty categories (not a problem here). (2) An intuitive p-score for each unit in your analysis is its fitted value from the logistic regression model (you might also use the exponentiated fitted value, the inverse of the fitted value, or other quantities depending on the circumstances). Fitted values from the results of a regression model in R (e.g., using the output of the command above) can be accessed using `model$fitted.values`.

```
# Alright, let's start by creating a formula for the p-score model:

f.pscore <- as.formula(paste("abd ~", paste(covars, collapse = "+") ) )
pscore.model <- glm(f.pscore, data = d, family = binomial(link =
                                          logit) )
pscore <- pscore.model$fitted.values
summary(pscore.model)

##
## Call:
## glm(formula = f.pscore, family = binomial(link = logit), data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9148  -1.1811   0.7252   0.9406   1.7903
```

```
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.05126    0.54092   0.095  0.92450
## C_ach        0.40425    0.30908   1.308  0.19090
## C_akw        1.08897    0.33908   3.212  0.00132 **
## C_ata       -0.58620    0.30170  -1.943  0.05202 .
## C_kma        0.48642    0.31118   1.563  0.11802
## C_oro       -0.74685    0.34511  -2.164  0.03046 *
## C_pad        0.23175    0.31458   0.737  0.46131
## C_paj        0.54579    0.31490   1.733  0.08306 .
## C_pal             NA         NA      NA       NA
## age          0.05057    0.01646   3.072  0.00212 **
## fthr_ed     -0.01813    0.02419  -0.749  0.45367
## mthr_ed     -0.03483    0.02870  -1.213  0.22503
## orphan96    -0.08751    0.30179  -0.290  0.77183
## hh_fthr_frm -0.13214    0.28303  -0.467  0.64059
## hh_size96   -0.05692    0.02047  -2.781  0.00542 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 981.58  on 740  degrees of freedom
## Residual deviance: 917.51  on 727  degrees of freedom
## AIC: 945.51
##
## Number of Fisher Scoring iterations: 4
```

Part b Assess balance on p-scores across treated and untreated units. Inlcude any descriptive statistics you think are relevant as well as the results of any statistical tests you use in this assessment.

```
# Descriptive differences first:
summary(pscore[d$abd == 0])

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2477  0.4495  0.5914  0.5702  0.6845  0.8401

summary(pscore[d$abd == 1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2014  0.5909  0.6749  0.6557  0.7485  0.9025
```

```
t.test(pscore[d$abd==1], pscore[d$abd == 0])
```

```
##
##  Welch Two Sample t-test
##
## data:  pscore[d$abd == 1] and pscore[d$abd == 0]
## t = 8.0355, df = 520.524, p-value = 6.287e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.06456393 0.10634906
## sample estimates:
## mean of x mean of y
## 0.6556577 0.5702012
```

```
ks.test(pscore[d$abd == 1], pscore[d$abd == 0])
```

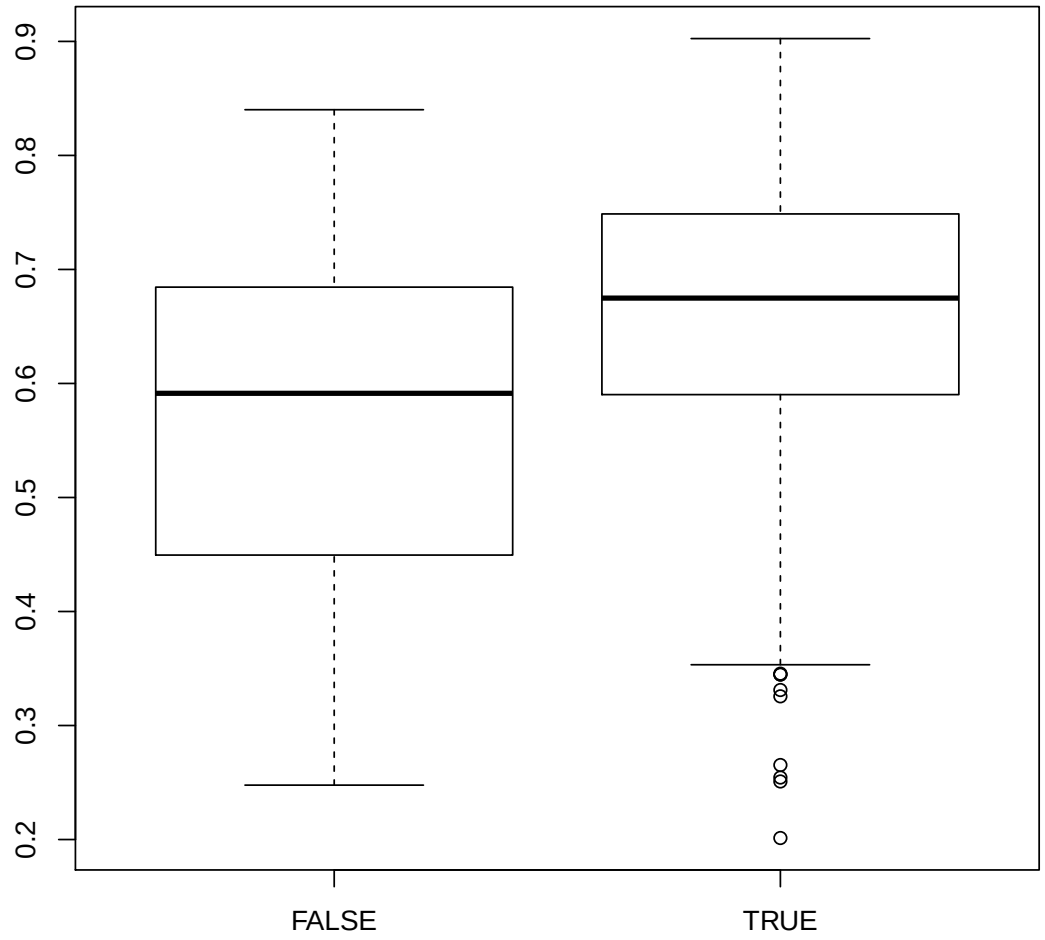
```
## Warning in ks.test(pscore[d$abd == 1], pscore[d$abd == 0]):  p-value will
be approximate in the presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  pscore[d$abd == 1] and pscore[d$abd == 0]
## D = 0.2713, p-value = 1.525e-11
## alternative hypothesis: two-sided
```

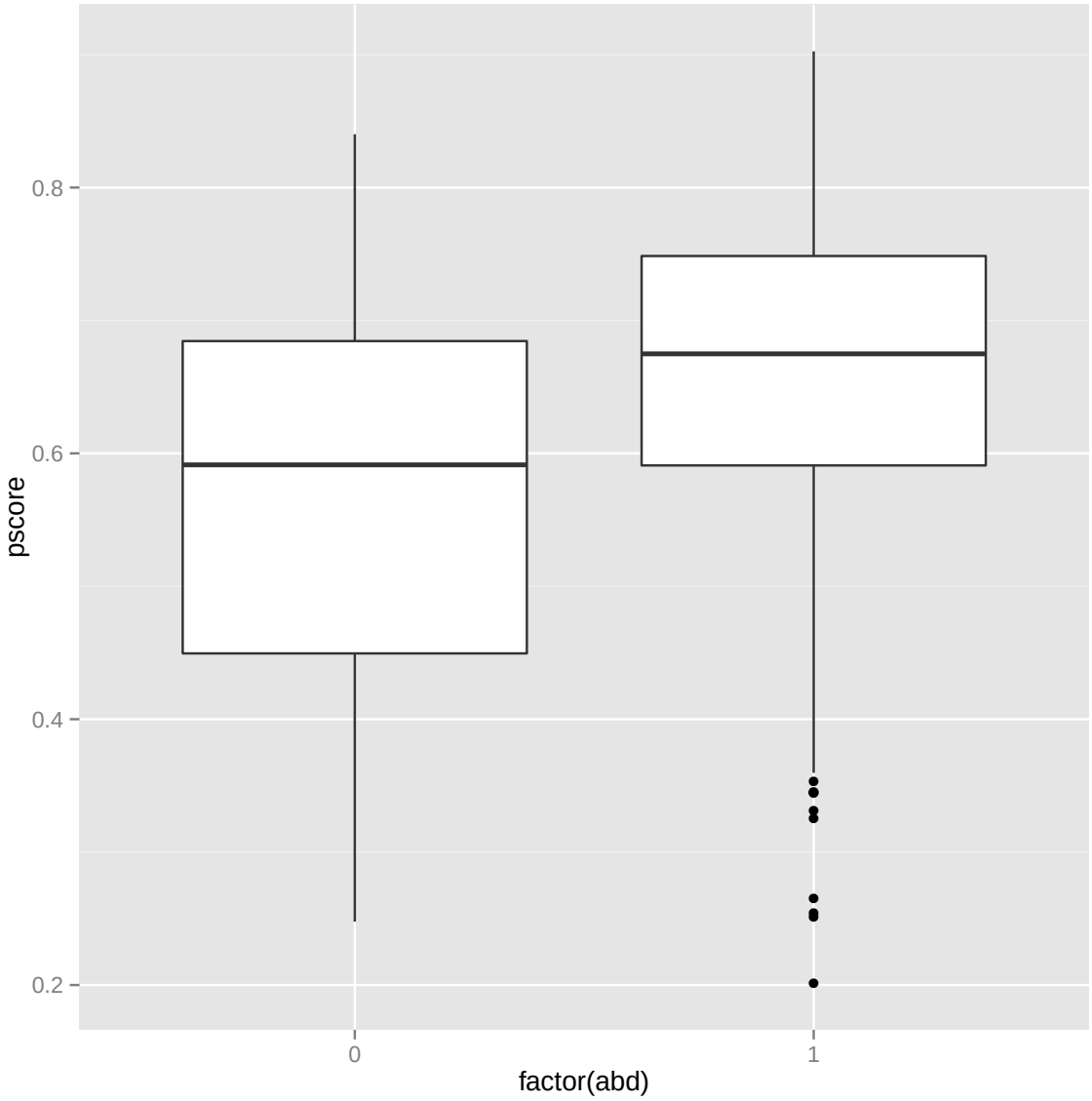
> Not surprisingly (based on the earlier balance assessments), the propensity score distribution in treatment and control groups is imbalanced. In other words, some combinations of observed attributes made it extremely likely that individuals were either going to be abducted (or not).

Part c Visually assess balance on the propensity score with a box-plot or a density plot. Before you run the command, consider: what do you expect the graph(s) to look like?

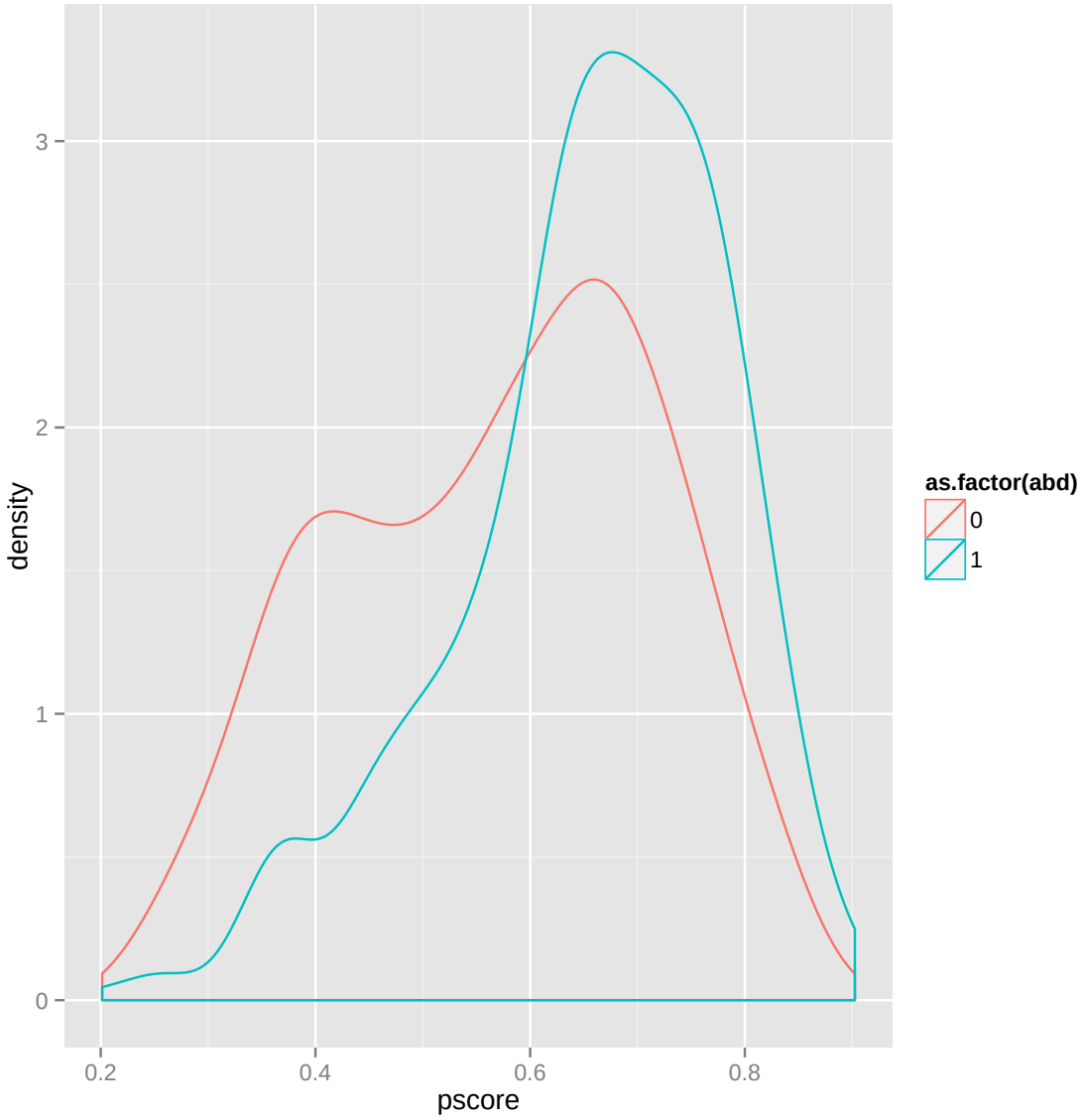
```
boxplot(pscore ~ as.logical(d$abd))
```




```
ggplot(d, aes(factor(abd), pscore)) + geom_boxplot()
```




```
ggplot(d, aes(pscore, colour=as.factor(abd))) + geom_density()
```



Notice how the biggest difference occur at the extremes? Some units in the treatment condition have p-scores very close to 1 (they were almost 100% likely to receive treatment), whereas some units in control have p-scores very close to 0 (they had almost no chance of receiving treatment). It does not make sense to estimate treatment effects at these extreme levels of the propensity score because they are outside the "area of common support" — the range within which subjects actually had an approximately equal ("as-if random") probability of being assigned to treatment *or* control.

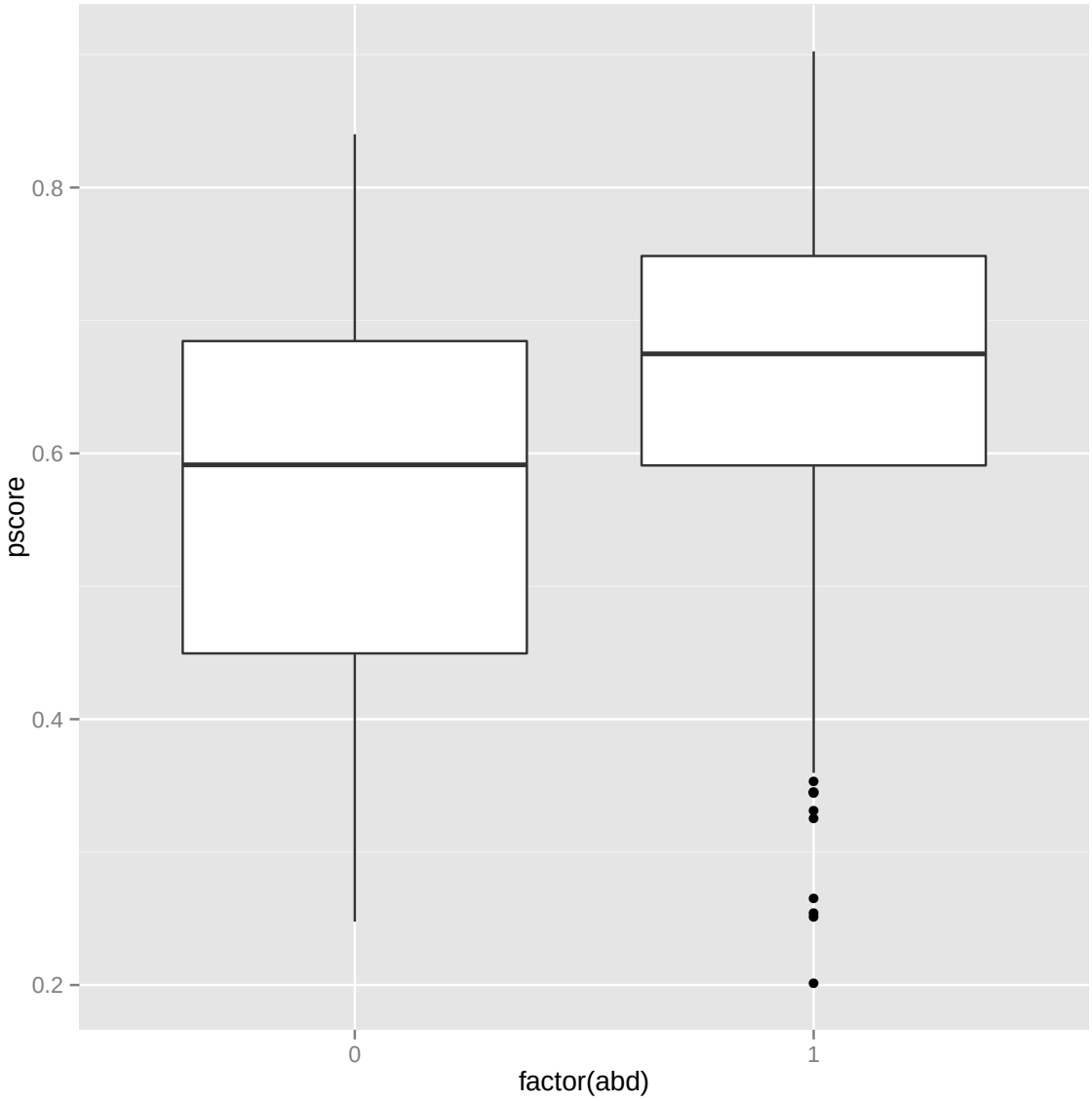
Part d Trim the data set so that only units with propensity score values between the 10^{th} — 90^{th} percentiles of all propensity scores remain. You'll want to use the `quantile()`

command to locate the values of the p-scores corresponding to these percentiles.

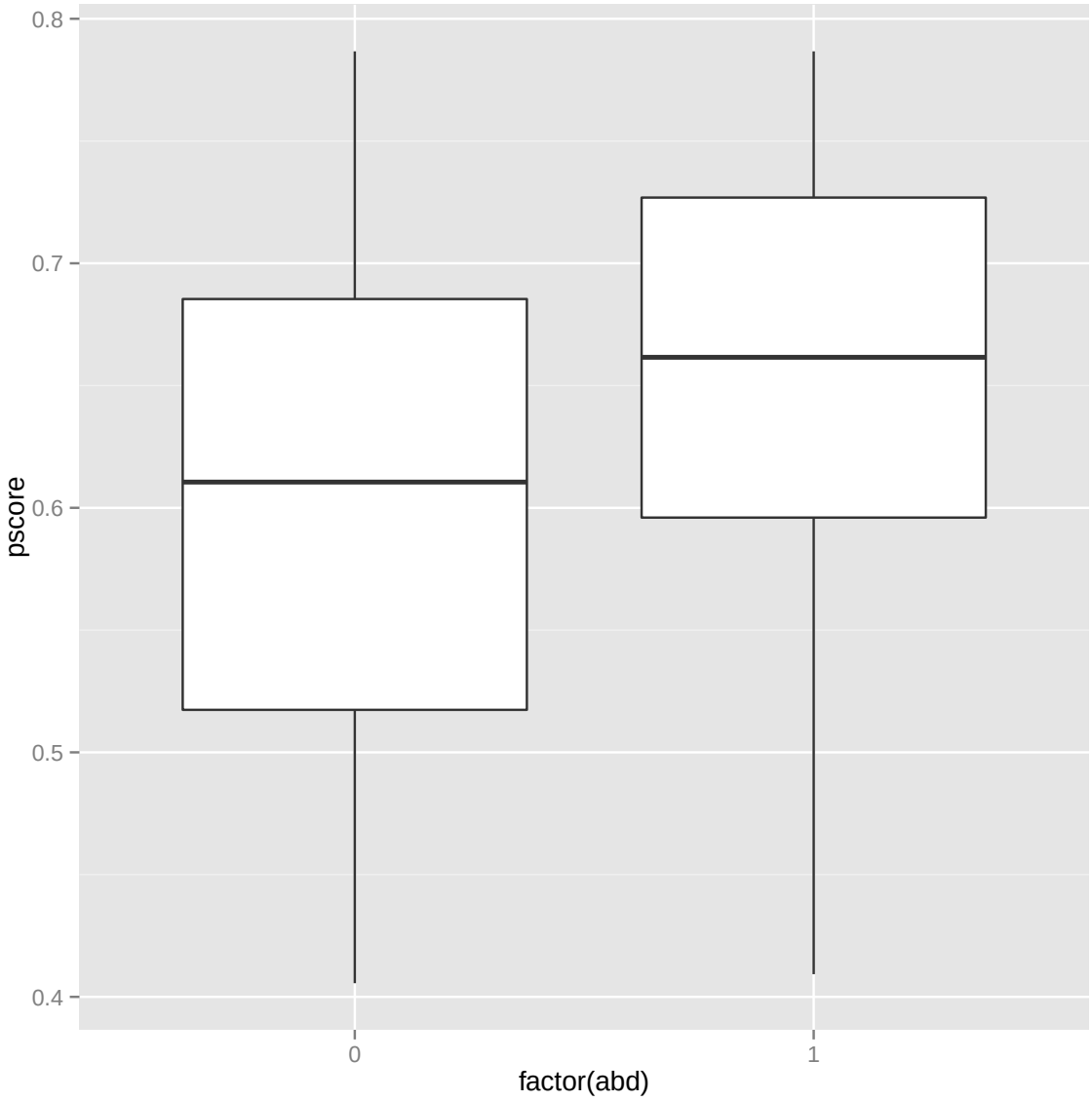
```
# first, we'll need to generate the decile values that we can
# then use for trimming the dataset
deciles <- quantile(pscore, c(.1,.9))

# take a look at them to see what we've got:
deciles

##       10%       90%
## 0.4012783 0.7866817

# add pscores to the dataset and trim the data at .1, .9 thresholds
d$pscore <- pscore
trim.d <- d[d$pscore > deciles[1] & d$pscore <= deciles[2],]
```

Part e Assess the covariate balance between treatment and control groups (use the original covariates, not the p-scores). Be sure to conduct any statistical and visual comparisons you deem relevant. Compare your results here to the results of Question 1, Part c. Discuss what you find in this comparison. What sorts of units did the trimming procedure remove from the dataset?

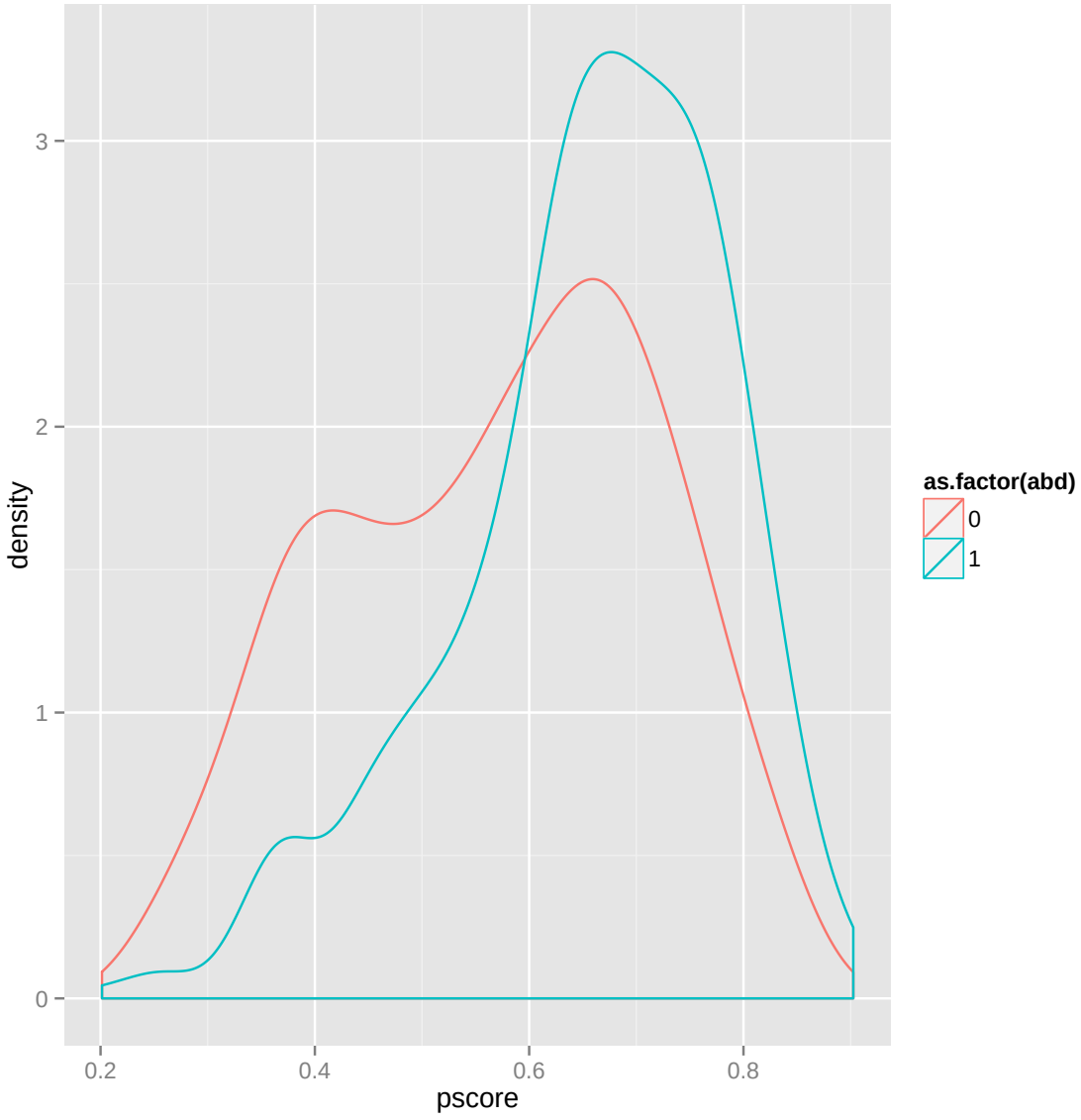
```
# compare balance using a boxplot
ggplot(d, aes(factor(abd), pscore)) + geom_boxplot() # old balance
```



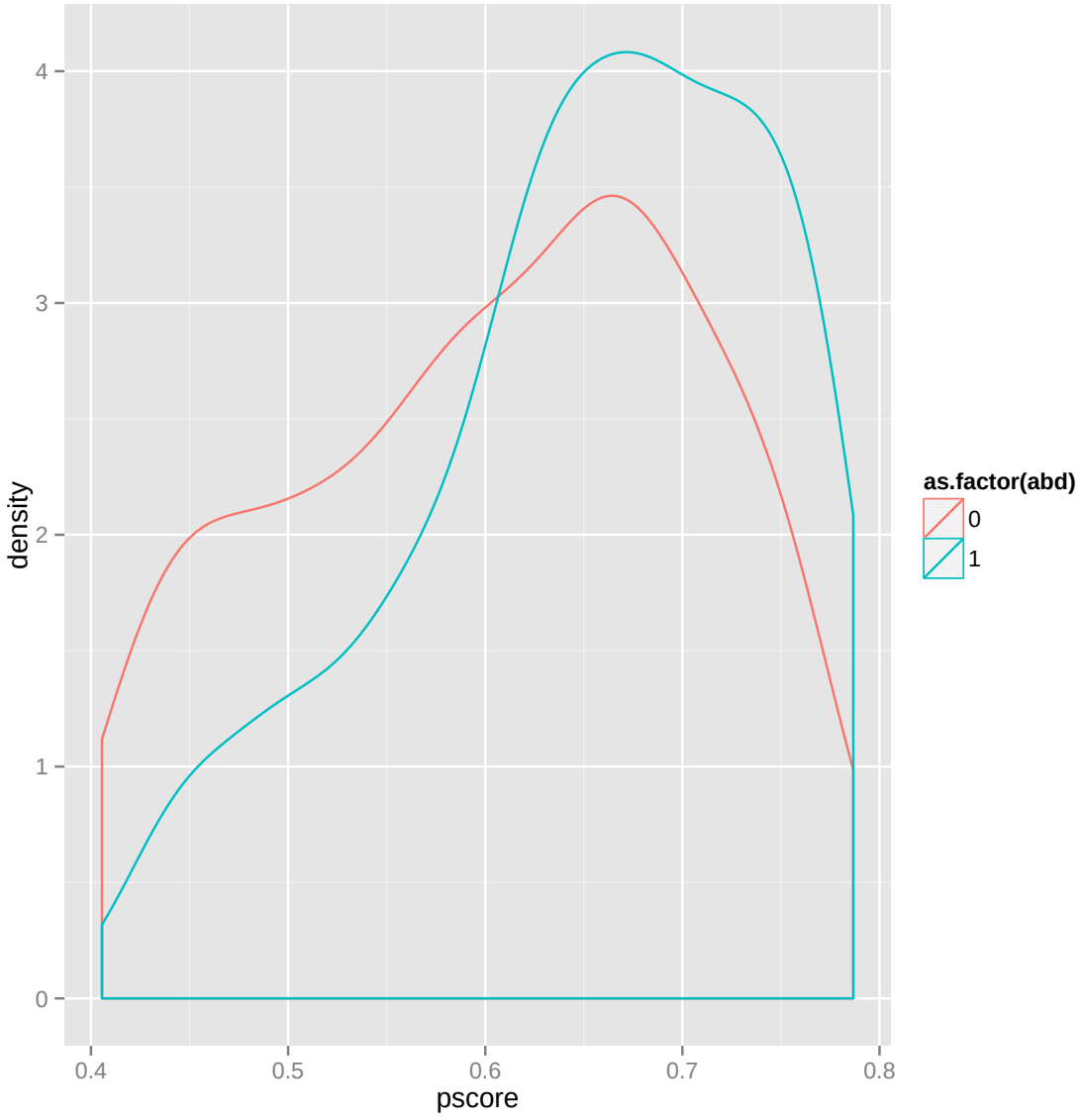

```
ggplot(trim.d, aes(factor(abd), pscore)) + geom_boxplot() # new balance
```




```
# and a density plot
ggplot(d, aes(pscore, colour=as.factor(abd))) + geom_density() # old balance
```




```
ggplot(trim.d, aes(pscore, colour=as.factor(abd))) + geom_density() # new balance
```



Part f Estimate the average treatment effect (ATE) of abduction on educational attainment using a linear regression model with inverse propensity score weights applied to all of the units. (Think: why does it make sense to weight using the inverse of the propensity score?) Create the weights as a new variable equal to the inverse of the propensity score for all units in the treatment condition and the inverse of 1 minus the propensity score for all units in the control condition. When you run the regression, you can pass this weights variable as an additional argument to the `lm()` or `glm()` command (e.g., `weights = d$my.weights`).

```r
# generate inverse propensity weights
trim.d$w <- trim.d$abd/trim.d$pscore +((1-trim.d$abd)/(1-trim.d$pscore))

# and now estimate treatment effect within the .1, .9 interval of the
# pscore passing in the weights as an additional argument to the lm()
# function:

# first without covariate adjustment:
summary(glm(educ ~ abd, data = trim.d, weights = trim.d$w))

##
## Call:
## glm(formula = educ ~ abd, data = trim.d, weights = trim.d$w)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -14.9296   -2.4718   -0.8705    2.6074   17.1176
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.6498     0.1768  43.272  < 2e-16 ***
## abd          -0.7298     0.2479  -2.944  0.00337 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 18.03745)
##
##     Null deviance: 10798  on 591  degrees of freedom
## Residual deviance: 10642  on 590  degrees of freedom
## AIC: 3024.1
##
## Number of Fisher Scoring iterations: 2

# and now with the fully specified model:
summary(glm(f.educ, data = trim.d, weights = trim.d$w))

##
## Call:
## glm(formula = f.educ, data = trim.d, weights = trim.d$w)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -14.2364   -2.5304   -0.4174    2.5229   14.4236
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.85280    0.79552   7.357 6.49e-13 ***
## abd         -0.74609    0.23380  -3.191  0.00149 **
## C_ach       -1.92513    0.42817  -4.496 8.37e-06 ***
## C_akw       -0.83006    0.53401  -1.554  0.12064
## C_ata       -1.35435    0.45936  -2.948  0.00332 **
## C_kma       -0.76722    0.43168  -1.777  0.07605 .
## C_oro       -2.08533    0.69670  -2.993  0.00288 **
## C_pad       -1.33248    0.43664  -3.052  0.00238 **
## C_paj       -0.66395    0.43928  -1.511  0.13122
## C_pal             NA         NA      NA       NA
## age          0.05785    0.02655   2.179  0.02972 *
## fthr_ed      0.16968    0.03515   4.828 1.77e-06 ***
## mthr_ed      0.06578    0.04246   1.549  0.12186
## orphan96     0.77063    0.46004   1.675  0.09445 .
## hh_fthr_frm -0.40611    0.40669  -0.999  0.31842
## hh_size96    0.09155    0.03348   2.734  0.00644 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 15.82419)
##
##     Null deviance: 10798.4  on 591  degrees of freedom
## Residual deviance:  9130.6  on 577  degrees of freedom
## AIC: 2959.4
##
## Number of Fisher Scoring iterations: 2
```

Part g Discuss the results of Question 2 Part e. Consider the differences between your estimates after propensity-score weighting with the results of "naive regression" in Question 1, Part a. Which of these estimates do you (not) find to be credible? Why? What are the limitations, if any, of your analysis given the particular characteristics of the research design and the balancing procedure you have pursued here?

> The analysis of the trimmed and inverse propensity weighted data suggest that abduction into the LRA caused victims to experience approximately 0.75 years

less schooling (std. err. = 0.23). This estimate is robust to the covariate adjustment.

Comparing these results with the naive estimates (with and without covariate adjustment), we can see that the unadjusted estimate in Question 1 Part a was likely different due to bias introduced by the imbalance between treatment and control groups. Although the covariate-adjusted "naive" estimate was closer to the result of the trimmed and weighted data, the latter procedure produces results that are preferable because they (a) are adjusted to account for variation in the probability of individuals with certain attributes to receive treatment or not (by virtue of the inverse propensity score weighting); (b) have only incorporated data from the region of common support (by virtue of the trimming); and (c) likely have enhanced precision due to covariance adjustment. The estimate generated in Part f is thus an estimate of a Local Average Treatment Effect (LATE), restricting our ability to generalize more broadly to the full (untrimmed) sample (nevermind the population as a whole). However, given the particular nature of the empirical circumstances under analysis it's not clear that broader generalizability beyond the individuals "at-risk" of abduction would have been possible anyway.