POLS 207

Problem Set 3*

Villaseñor-Derbez, J.C.

Problem 1: NSW experiment and observational comparison

```
# Load packages
suppressPackageStartupMessages({
  library(startR)
  library(Matching)
  library(ebal)
  library(foreign)
  library(here)
  library(tidyverse)
})

# Load data
nsw_dat <- read.dta("nsw_exper.dta")
```

Using the experimental data, obtain a simple unbiased estimate of the ATE of NSW on 1978 earnings, its SE and a 95% confidence interval.

Because this is experimental data, the difference in means is an unbiased estimate of the ATE.

```
# Simple unbiased estimate of the ATE
mT <- mean(nsw_dat$re78[nsw_dat$nsw == 1], na.rm = T)
mC <- mean(nsw_dat$re78[nsw_dat$nsw == 0], na.rm = T)

ATE <- mT - mC
```

```
# Calculate standard errors
# Get variances
sigma_y1 <- var(nsw_dat$re78[nsw_dat$nsw == 1], na.rm = T)
sigma_y0 <- var(nsw_dat$re78[nsw_dat$nsw == 0], na.rm = T)

# Get sample sizes
N1 <- sum(nsw_dat$nsw == 1, na.rm = T)
N0 <- sum(nsw_dat$nsw == 0, na.rm = T)

# Calculate Standard Errors
SE <- sqrt((sigma_y1 / N1) + (sigma_y0 / N0))
```

```
# Calculate 95% CIs
CI_h <- ATE + (1.96 * SE)
CI_l <- ATE - (1.96 * SE)
```

The $A\hat{T}E = 1794.343$, with $SE = 670.99$ and $CI_{95} = (479.18, 3109.49)$.

[*] Available on GitHUb: https://github.com/jcvdav/POLS207/blob/master/ps3/ps3.pdf

With the experimental data, use OLS to estimate the ATE controlling for age, education, race, ethnicity, marital status, and employment and earnings in 1974 and 1975. Report the estimate and its SE. Compare it to the one obtained in (a), explaining the reason for any similarities or differences.

The table below shows the coefficient estimates for the stated regression. The $A\hat{T}E_{OLS} = 1682.58$. The previous estimate was slightly higher ($A\hat{T}E = 1794.343$), but the OLS estimate is within the 95% CIs calculated before.

```r
lm(re78 ~ nsw + age + educ + black + hisp + married + re74 + re75 + u74 + u75, data = nsw_dat) %>%
  stargazer::stargazer(.,
                       se = estimatr::starprep(.,
                                               se_type = "HC2"),
                       t.auto = T,
                       p.auto = T,
                       header = F,
                       title = "Estimate of the Sample Average Treatment Effect.",
                       single.row = T
  )
```

Table 1: Estimate of the Sample Average Treatment Effect.

	Dependent variable:
	re78
nsw	1,672.042** (663.722)
age	53.668 (40.567)
educ	402.947** (163.077)
black	−2,039.466* (1,048.241)
hisp	424.649 (1,443.157)
married	−146.662 (870.023)
re74	0.124 (0.133)
re75	0.019 (0.144)
u74	1,380.999 (1,571.072)
u75	−1,071.817 (1,411.349)
Constant	221.429 (2,864.634)
Observations	445
R^2	0.058
Adjusted R^2	0.037
Residual Std. Error	6,509.273 (df = 434)
F Statistic	2.683*** (df = 10; 434)
Note:	*p<0.1; **p<0.05; ***p<0.01

File `nsw_psid_withtreated.dta` contains treated units taken from the experiment, but control units replaced by the non-experimental sample from the PSID.

```r
psid_dat <- read.dta("nsw_psid_withtreated.dta")
```

Check the covariate balance in this merged dataset. Decide on a few sensible balance statistics and report them in a table.

```r
balance <- psid_dat %>%
  drop_na() %>%
  MatchBalance(
    nsw ~ age + educ + black + hisp + married + re74 + re75 + u74 + u75,
    data = .,
    print.level = 0
  ) %>%
  baltest.collect(
    var.names = c(
      "age",
      "educ",
      "black",
      "hisp",
      "married",
      "re74",
      "re75",
      "u74",
      "u75"),
    after = F
  ) %>%
  as_tibble(rownames = "Covariate") %>%
  dplyr::select(-contains("qq"), -contains("pooled"))

knitr::kable(
  balance,
  booktabs = T,
  col.names = c(
    "Covariate",
    "Mean (Treatment)",
    "Mean (Control)",
    "Standardized difference",
    "Variance ratio",
    "T p-value",
    "KS p-value"
  ),
  caption = "Pre-matching balance of covariates."
)
```

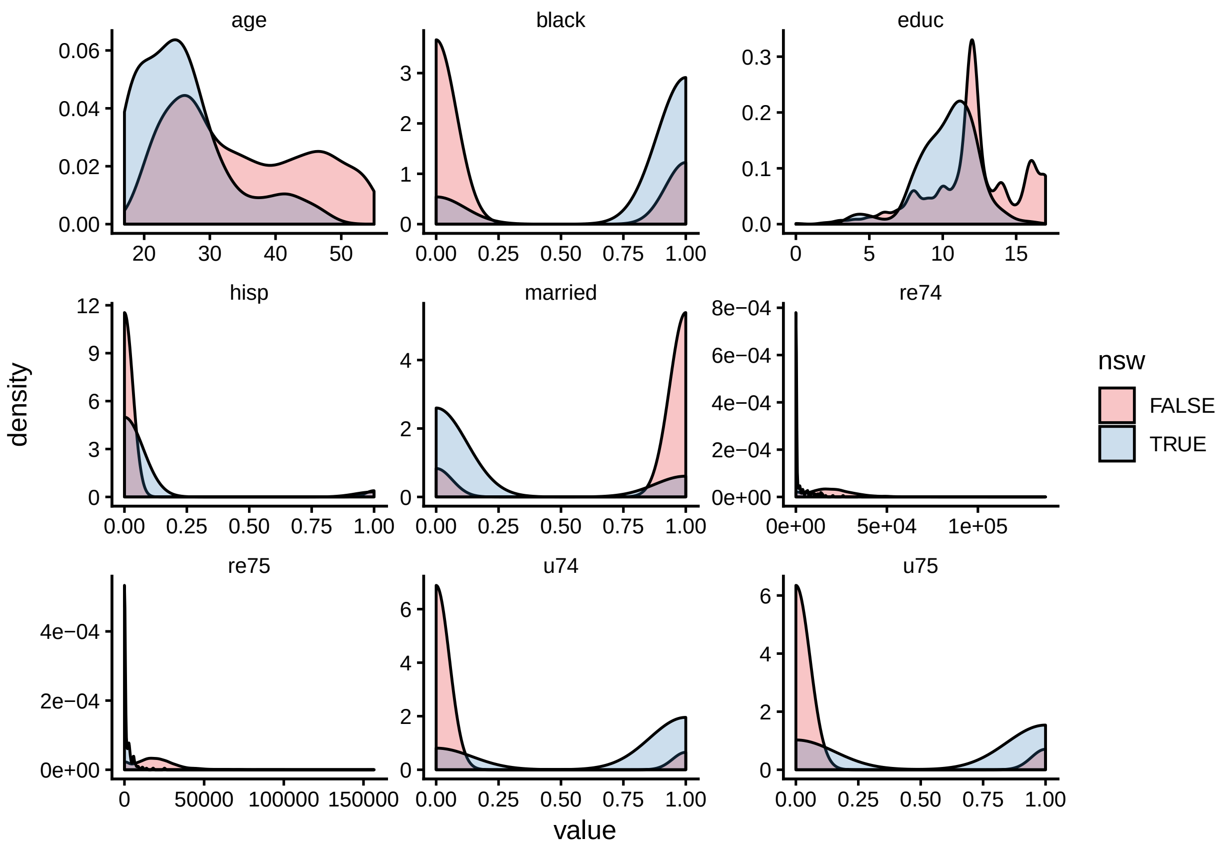
Table 2: Pre-matching balance of covariates.

Covariate	Mean (Treatment)	Mean (Control)	Standardized difference	Variance ratio	T p-value	KS p-value
age	25.8162162	3.485060e+01	-126.26641	0.4696319	0.0000000	0
educ	10.3459459	1.211687e+01	-88.07705	0.4254861	0.0000000	0
black	0.8432432	2.506024e-01	162.56425	0.7073935	0.0000000	NA
hisp	0.0594595	3.253010e-02	11.35663	1.7858904	0.1317327	NA
married	0.1891892	8.662651e-01	-172.40603	1.3307596	0.0000000	NA
re74	2095.5740112	1.942875e+04	-354.70658	0.1328502	0.0000000	0
re75	1532.0556149	1.906334e+04	-544.57642	0.0560566	0.0000000	0
u74	0.7081081	8.634540e-02	136.39138	2.6331760	0.0000000	NA
u75	0.6000000	1.000000e-01	101.78586	2.6800826	0.0000000	NA

How do the treatment and control group differ?

The treatment group has younger people, with less education, a greater percentage of black and hispanics, and most people are single. The treatment group has lower average real earnings. The figure below shows the density distributions for each covariate in the dataset.

```
psid_dat %>%
  select(-c(re78, u78)) %>%
  gather(variable, value, -c(nsw)) %>%
  mutate(nsw = nsw == 1) %>%
  ggplot(aes(x = value, fill = nsw, group = nsw)) +
  geom_density(alpha = 0.25) +
  facet_wrap(~variable, scales = "free") +
  ggtheme_plot() +
  scale_fill_brewer(palette = "Set1")
```



Among the observed covariates, what seem to be the most important factors that determine selection into the program (the "treatment")?

The data suggest that your real earnings for 1974 and 1975 and ethnicity are prediuctors of selection into the program.

Estimate the (naive) ATE of the program on 1978 earnings without adjusting for any of the covariates. Report the estimate and a standard error.

```r
# Simple unbiased estimate of the ATE for psid data
mTpsid <- mean(psid_dat$re78[psid_dat$nsw == 1], na.rm = T)
mCpsid <- mean(psid_dat$re78[psid_dat$nsw == 0], na.rm = T)

ATEpsid <- mTpsid - mCpsid
```

```r
# Calculate standard errors
# Get variances
sigma_y1psid <- var(psid_dat$re78[psid_dat$nsw == 1], na.rm = T)
sigma_y0psid <- var(psid_dat$re78[psid_dat$nsw == 0], na.rm = T)

# Get sample sizes
N1psid <- sum(psid_dat$nsw == 1, na.rm = T)
N0psid <- sum(psid_dat$nsw == 0, na.rm = T)

# Calculate Standard Errors
SEpsid <- sqrt((sigma_y1psid / N1) + (sigma_y0psid / N0))
```

In this case, the naive ATE estimate for 1978 earnigns appears to be $ATE = -15204.78$, suggesting that the program reduces the average earnings. The standard error is now SE1124.821

Repeat (b) using the non-experimental data. Does the estimate of the ATE change? Why or why not?

The table below shows a drastic change in my ATE. By including covariates, the sign and magnitude of the ATE changes. This, however, still produces an estimate lower to the one obtaines with experimental data.

```r
lm(re78 ~ nsw + age + educ + black + hisp + married + re74 + re75, data = psid_dat) %>%
  stargazer::stargazer(.,
                       se = estimatr::starprep(.,
                                               se_type = "HC2"),
                       t.auto = T,
                       p.auto = T,
                       header = F,
                       title = "Estimate of the Sample Average Treatment Effect using the non-experimen
                       single.row = T
  )
```

Using the non-experimental data, condition (only) on the marital status of individuals, and manually compute the subclassification estimator of the ATT.

```r
married_treated <- psid_dat$re78[psid_dat$nsw == 1 & psid_dat$married == 1]
married_control <- psid_dat$re78[psid_dat$nsw == 0 & psid_dat$married == 1]

single_treated <- psid_dat$re78[psid_dat$nsw == 1 & psid_dat$married == 0]
single_control <- psid_dat$re78[psid_dat$nsw == 0 & psid_dat$married == 0]

married_ATT <- mean(married_treated) - mean(married_control)
single_ATT <- mean(single_treated) - mean(single_control)
```

Table 3: Estimate of the Sample Average Treatment Effect using the non-experimental data.

	Dependent variable:
	re78
nsw	859.769 (768.537)
age	-81.537^{***} (20.719)
educ	528.024^{***} (88.743)
black	-542.706 (442.463)
hisp	$2{,}165.572^{*}$ (1,227.276)
married	$1{,}220.269^{**}$ (496.886)
re74	0.278^{***} (0.062)
re75	0.568^{***} (0.067)
Constant	776.729 (1,489.169)
Observations	2,675
R^2	0.586
Adjusted R^2	0.585
Residual Std. Error	10,070.410 (df = 2666)
F Statistic	472.194^{***} (df = 8; 2666)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
n_treated <- sum(psid_dat$nsw == 1)
n_treated_married <- sum(psid_dat$nsw == 1 & psid_dat$married == 1)
pr_married_treated <- n_treated_married / n_treated
pr_single_treated <- 1 - pr_married_treated

ATT_married <- (pr_single_treated * single_ATT) + (pr_married_treated * married_ATT)
```

The subclassification ATT conditioning on marital status is $ATT = -11124.4$.

Repeat the above, this time conditioning (only) on Unemployment status in 1975 using a sub-classification estimator of the ATT.

```r
unemployed_treated <- psid_dat$re78[psid_dat$nsw == 1 & psid_dat$u75 == 1]
unemployed_control <- psid_dat$re78[psid_dat$nsw == 0 & psid_dat$u75 == 1]

employed_treated <- psid_dat$re78[psid_dat$nsw == 1 & psid_dat$u75 == 0]
employed_control <- psid_dat$re78[psid_dat$nsw == 0 & psid_dat$u75 == 0]

unemployed_ATT <- mean(unemployed_treated) - mean(unemployed_control)
employed_ATT <- mean(employed_treated) - mean(employed_control)

n_treated_unemployed <- sum(psid_dat$nsw == 1 & psid_dat$u75 == 1)
pr_unemployed_treated <- n_treated_unemployed / n_treated
pr_employed_treated <- 1 - pr_unemployed_treated

ATT_u75 <- (pr_employed_treated * employed_ATT) + (pr_unemployed_treated * unemployed_ATT)
```

The subclassification ATT conditioning on employment status is $ATT = -6244.687$.

Problem 2: Matching on NSW

With the non-experimental data, show the balance on the data. Then match, using the following covariates: "age", "educ", "black", "hisp", "married", "re74", "re75", "u74", and "u75". Show the new balance tables, and estimate the ATT.

The table below shows post-matching balance, where balance is achieved for most of the covariates (not for education or re75). Estimating the ATT with biad adjustment, our ATT estimate is eroded ($ATT = 1684$; $t_{184} = 1.432; p = 0.29$)

```r
X <- psid_dat %>%
  select(-c(nsw, re78, u78)) %>%
  as.matrix()

matched <- Match(Y = psid_dat$re78,
                 Tr = psid_dat$nsw,
                 X = X, M = 1,
                 estimand = "ATT",
                 BiasAdjust = T)

balance_matched <- MatchBalance(
  match.out = matched,
  formul = nsw ~ age + educ + black + hisp + married + re74 + re75 + u74 + u75,
  data = psid_dat,
  print.level = 0) %>%
  baltest.collect(
    var.names = c(
      "age",
      "educ",
      "black",
      "hisp",
      "married",
      "re74",
      "re75",
      "u74",
      "u75"),
    after = T
  ) %>%
  as_tibble(rownames = "Covariate") %>%
  dplyr::select(-contains("qq"), -contains("pooled"))

knitr::kable(
  balance_matched,
  booktabs = T,
  col.names = c(
    "Covariate",
    "Mean (Treatment)",
    "Mean (Control)",
    "Standardized difference",
    "Variance ratio",
    "T p-value",
    "KS p-value"
  ),
  caption = "Pre-matching balance of covariates."
)
```

Table 4: Pre-matching balance of covariates.

Covariate	Mean (Treatment)	Mean (Control)	Standardized difference	Variance ratio	T p-value	KS p-value
age	25.8162162	26.0774775	-3.651440	0.9240888	0.6010652	0.000
educ	10.3459459	10.6522523	-15.234191	1.3120104	0.0024980	0.000
black	0.8432432	0.8432432	0.000000	1.0000000	1.0000000	NA
hisp	0.0594595	0.0594595	0.000000	1.0000000	1.0000000	NA
married	0.1891892	0.1891892	0.000000	1.0000000	1.0000000	NA
re74	2095.5740112	2173.7482894	-1.599761	0.9734941	0.4333570	0.452
re75	1532.0556149	2095.3161514	-17.496633	0.7448871	0.0010229	0.052
u74	0.7081081	0.7081081	0.000000	1.0000000	1.0000000	NA
u75	0.6000000	0.6000000	0.000000	1.0000000	1.0000000	NA

What is the importance of the bias adjustment? When is it most important to use the bias adjustment?

In the non-experimental sample, compare the number of treated to untreated units. Comment on the result, and whether it is good or bad in your view

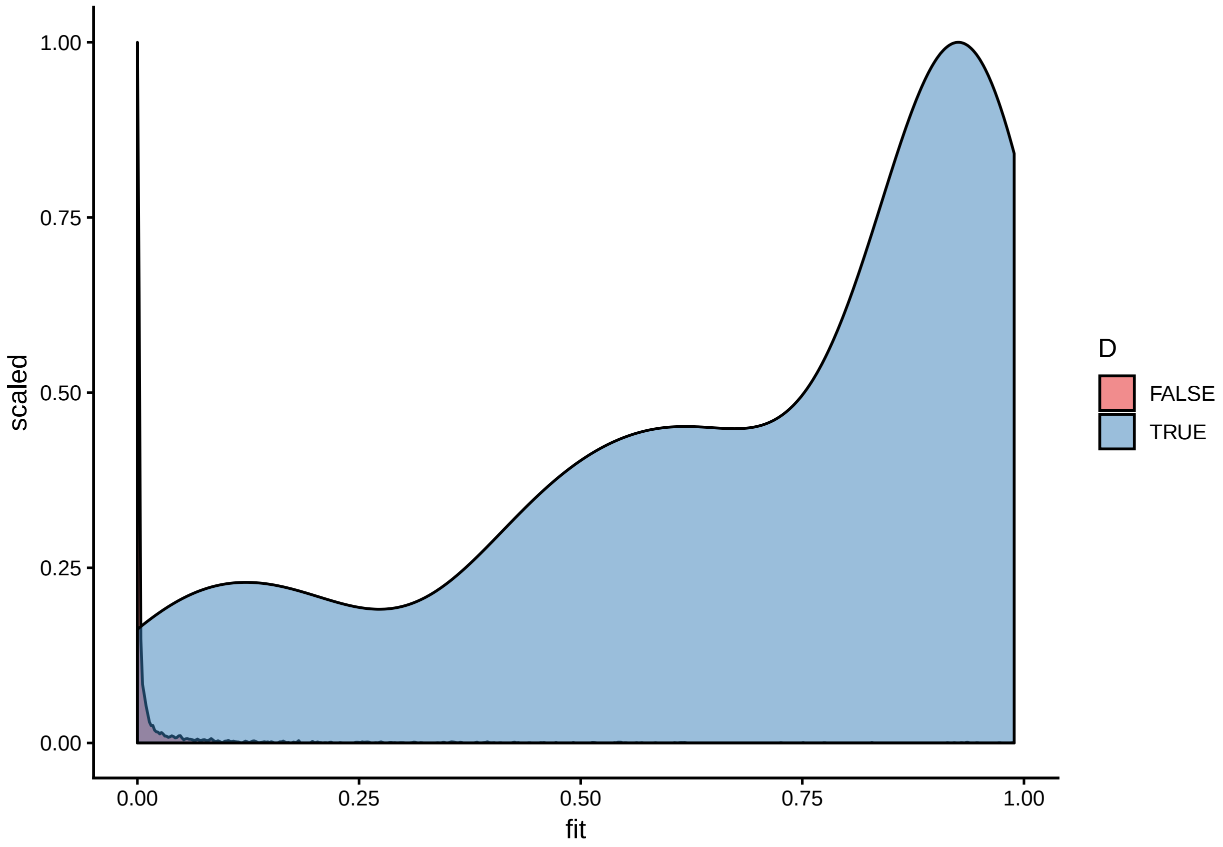
In the non-experimental sample, estimate propensity scores using a logistic regression. Report the distributions of propensity scores for treated and control groups and comment on the overlap.

```
glm_fit <- glm(nsw ~ age + educ + black + hisp + married + re74 + re75 + u74 + u75,
               data = psid_dat,
               family = "binomial" (link=logit))

fit <- glm_fit$fit

D <- psid_dat$nsw == 1
#
# PrD <- mean(D)
#
# IPW <- (D * PrD + (1-D) * (1-PrD)) / (D * fit + (1-D) * (1-fit))

tibble(D = D, fit = fit) %>%
  ggplot(aes(x  = fit, y = ..scaled.., group = D, fill = D)) +
  geom_density(alpha = 0.5) +
  ggtheme_plot() +
  scale_fill_brewer(palette = "Set1")
```



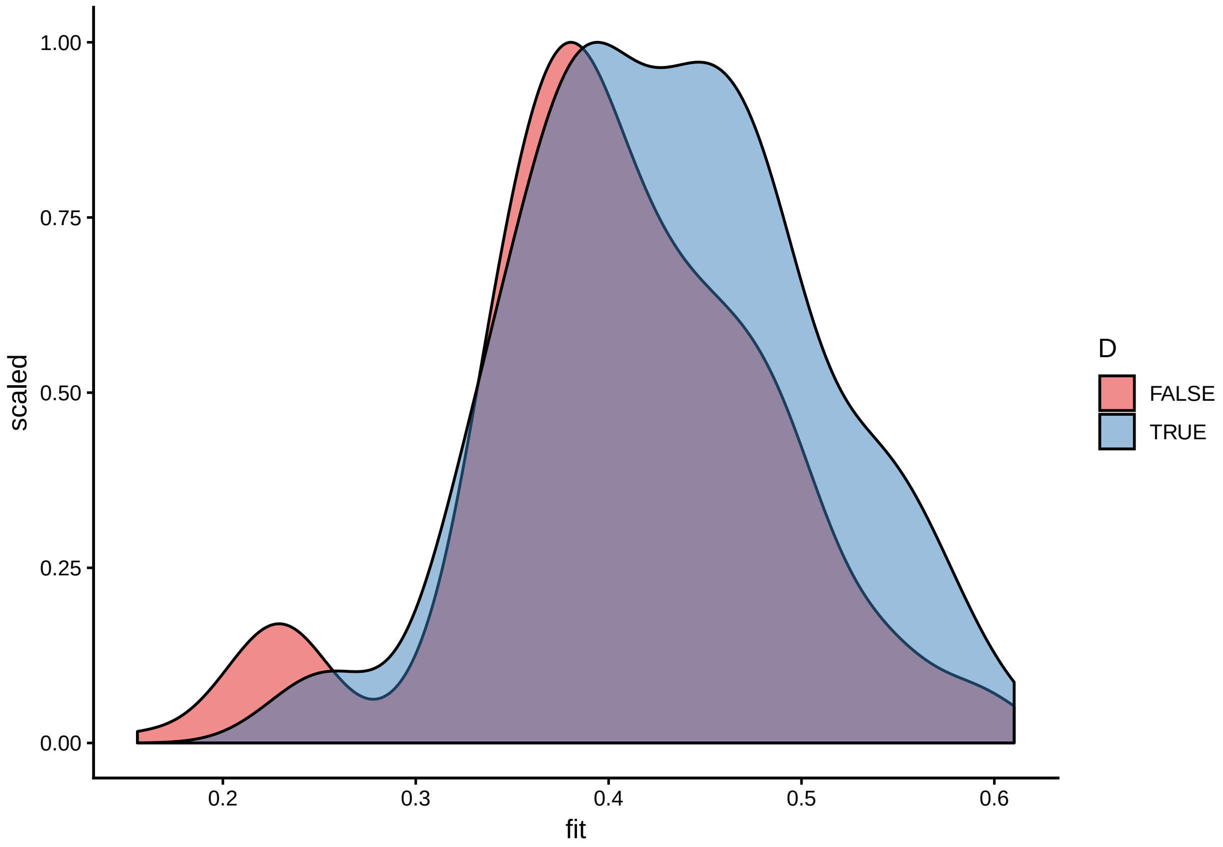
Now using the experimental sample, estimate propensity scores. Compare distributions of the propensity scores for treated and control groups here. What do you observe? Compare your results with part (d).

```
glm_fit_exp <- glm(nsw ~ age + educ + black + hisp + married + re74 + re75 + u74 + u75,
               data = nsw_dat,
               family = "binomial" (link=logit))

fit_exp <- glm_fit_exp$fit

D_exp <- nsw_dat$nsw == 1

tibble(D = D_exp, fit = fit_exp) %>%
  ggplot(aes(x  = fit, y = ..scaled.., group = D, fill = D)) +
  geom_density(alpha = 0.5) +
  ggtheme_plot() +
  scale_fill_brewer(palette = "Set1")
```


1.00
0.75
0.50
0.25
0.00
scaled
0.2
0.3
0.4
0.5
0.6
fit
D
🟥 FALSE
🟦 TRUE