STAT 528 - Advanced Regression Analysis II

Binary response regression (part II)

Daniel J. Eck
Department of Statistics
University of Illinois

Last time

- logistic regression
- data analysis
- connecting theory to application

Learning Objectives Today

- basic diagnostics
- probit regression and threshold modeling
- basic causal inference

Example: CCSO data

We are continuing with our demonstration of logistic regression modeling on the Champaign County Sheriff's Office (CCSO) data frame.

We now load in the CCSO data frame using the fread (**f**ast **read**) function from the `data.table` package (can also use read_csv in the `tidyverse`) and perform most of our data wrangling operations using `dplyr`.

```
rm(list = ls())
library(tidyverse)
library(data.table)
```

Let's load in and wrangle the data as before.

```
## load in data
system.time(CCSO <- fread("https://uofi.box.com/shared/static/9elozjsg99bgcb7gb546wlfr3r2gc9b7.csv"))
```

```
##    user  system elapsed
##   0.270   0.099   8.230
```

```
dim(CCSO)
```

```
## [1] 67764    35
```

```
## data wrangling
CCSO_small <- CCSO %>% rename(Days = "Days in Jail", Age = "Age at Arrest",
                          Date = "BOOKING DATE", Sex = "SEX", Race = "RACE",
                          Crime = "CRIME CODE", Agency = "ARREST AGENCY") %>%
  mutate(atleastone = ifelse(Days > 0,1,0)) %>%
  filter(Crime == "OTHER TRAFFIC OFFENSES") %>%
  filter(Race %in% c("Asian/Pacific Islander","Black","White","Hispanic")) %>%
  filter(Sex %in% c("Female","Male")) %>%
  dplyr::select(atleastone, Age, Sex, Date, Race) %>%
  mutate(Race = fct_drop(Race), Sex = fct_drop(Sex))
CCSO_small <- CCSO_small[complete.cases(CCSO_small), ]

head(CCSO_small, 5)
```

```
##    atleastone Age    Sex     Date     Race
## 1:          0  22   Male 1/1/2011    White
## 2:          0  26   Male 1/1/2011    White
## 3:          0  32 Female 1/1/2011    White
## 4:          0  22   Male 1/2/2011    White
## 5:          0  35   Male 1/2/2011 Hispanic
```

```
dim(CCSO_small)
```

```
## [1] 5916    5
```

We will continue on with our main effects only model.

```
m1 <- glm(atleastone ~ -1 + Race + Sex + Age, data = CCSO_small,
          family = "binomial", x = "TRUE")
p1 <- predict(m1, type = "response", se.fit = TRUE)

summary(m1)
```

```
##
## Call:
## glm(formula = atleastone ~ -1 + Race + Sex + Age, family = "binomial",
##     data = CCSO_small, x = "TRUE")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9393  -0.5485  -0.4817  -0.3391   2.6527
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## RaceAsian/Pacific Islander -4.389865   0.523612  -8.384  < 2e-16 ***
## RaceBlack                  -1.876550   0.144601 -12.977  < 2e-16 ***
## RaceHispanic               -2.804549   0.173349 -16.179  < 2e-16 ***
## RaceWhite                  -3.043226   0.147160 -20.680  < 2e-16 ***
## SexMale                     0.739834   0.105380   7.021 2.21e-12 ***
## Age                         0.007705   0.003186   2.418   0.0156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8201.3  on 5916  degrees of freedom
## Residual deviance: 4668.7  on 5910  degrees of freedom
## AIC: 4680.7
##
## Number of Fisher Scoring iterations: 6
```

Diagnostics

Model diagnostics for logistic regression fits are not as prevalent as those in linear regression.

We will discuss a straightforward binary response regression diagnostic method (Esarey and Pierce (2012)). More comprehensive diagnostics for GLMs will be discussed later.

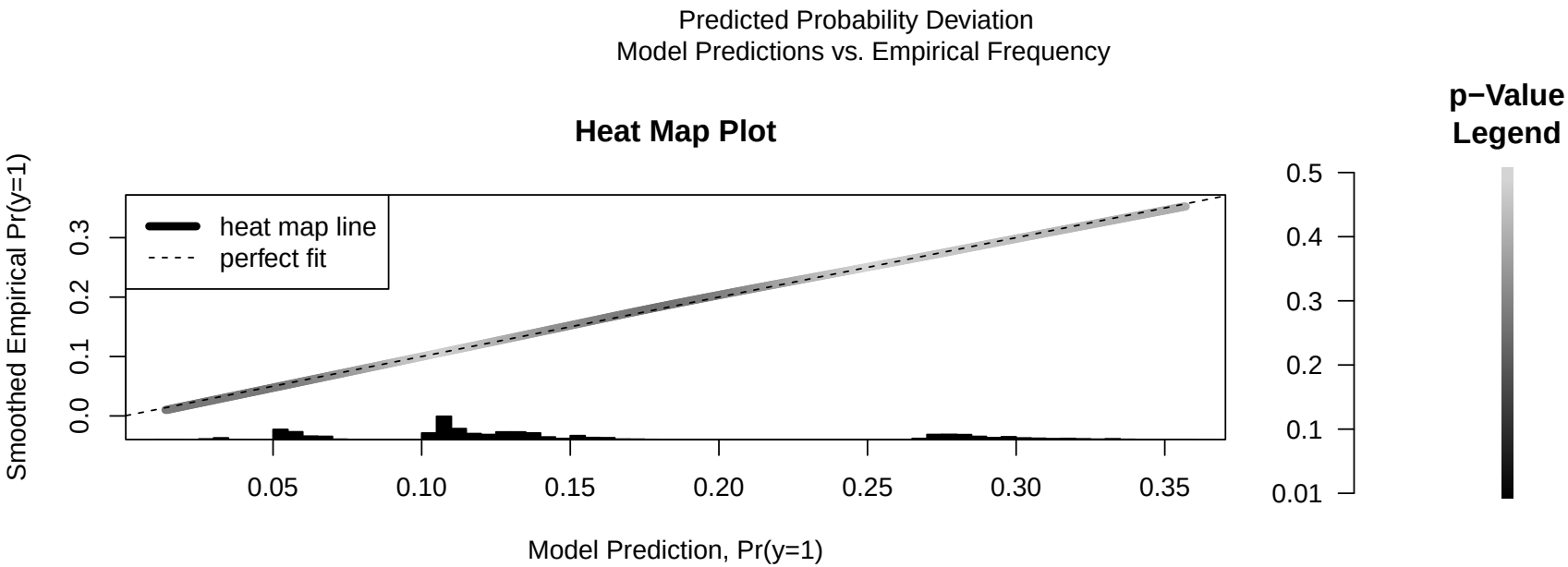
This method compares a model's predicted probability $\mathbb{P}(Y = 1)$ to the observed frequency of $Y = 1$ in the data set.

If the model is a good fit to the data, subsets of the data with $\mathbb{P}(Y = 1) \approx m$ should have about m proportion of cases for which $Y = 1$. This is the basis of Esarey and Pierce's method.

Esarey and Pierce's method is implemented in the `heatmap.fit` function within the `heatmapFit` package.

```
library(heatmapFit)
y <- CCSO_small$atleastone
heatmap.fit(y, p1$fit)
```

```
##
## Calculating optimal loess bandwith...
## aicc Chosen Span =  0.9899469
##
## Generating Bootstrap Predictions...
##   |                                                                      |
```



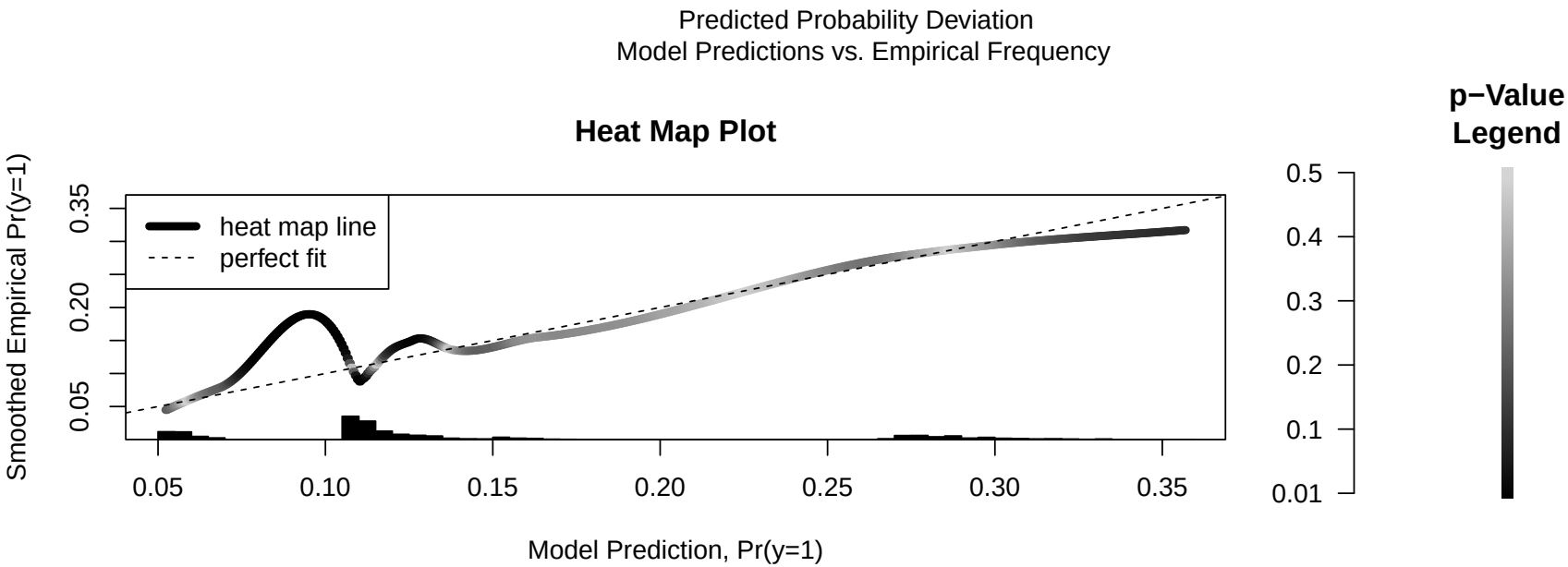

```
##
##
## ******************************************
## 0% of Observations have one-tailed p-value <= 0.1
## Expected Maximum = 20%
## ******************************************
```

Recall our smaller model that considers whether an individual was black or not.

```
CCSO_small <- CCSO_small %>% mutate(isBlack = ifelse(Race == "Black",1,0))
m2 <- glm(atleastone ~ isBlack + Sex + Age, data = CCSO_small,
          family = "binomial", x = "TRUE")
anova(m1, m2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: atleastone ~ -1 + Race + Sex + Age
## Model 2: atleastone ~ isBlack + Sex + Age
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      5910     4668.7
## 2      5912     4684.1 -2  -15.351 0.0004641 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Calculating optimal loess bandwith...
## aicc Chosen Span =  0.3441159
##
## Generating Bootstrap Predictions...
##   |                                                                      |
```




```
##
##
## ******************************************
## 29.68222% of Observations have one-tailed p-value <= 0.1
## Expected Maximum = 20%
## ******************************************
```

Probit regression

In this class we have focused on exponential theory, canonical links, and generalized linear models that arise from this perspective.

We now present an alternative to this paradigm in the form of the probit link function (inverse normal cdf) for binary response variables.

Probit regression arises from normal latent variable models.

Recall that we are interested in modeling $\pi(x) = \mathbb{P}(Y = 1|X = x)$ in the form

$$g^{-1}(\pi(x)) = x'\beta$$

where g^{-1} is not the logit function.

In probit regression we specify $g = \Phi$ where Φ is the CDF of a standard normal distribution so that

$$\Phi^{-1}(\pi(x)) = x'\beta.$$

A related latent variable approach, referred to as a threshold model or liability-threshold model, also specifies the probit model (See Wright (1934).

This model assumes that there is an unobserved continuous response y^* such that the observed response $y = 0$ if $y^* \leq \tau$ and $y = 1$ if $y^* > \tau$.

Suppose that $y^* = \mu + \varepsilon$, where $\mu = x'\beta$ and where ε_i are independent realizations from a $N(0, \sigma^2)$ distribution. Then,

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid X = x) &= \mathbb{P}(Y^* > \tau \mid X = x) \\
&= \mathbb{P}(x'\beta + \varepsilon > \tau \mid X = x) \\
&= \mathbb{P}(-\varepsilon < x'\beta - \tau \mid X = x) \\
&= \Phi\left((x'\beta - \tau)/\sigma\right)
\end{aligned}$$

where it is noted that $\varepsilon \overset{d}{=} -\varepsilon$.

There is no information in the data about σ or the threshold τ. An equivalent model results if we multiply (β', σ, τ) by any positive constant.

For identifiability, we set $\sigma = 1$ and $\tau = 0$, and the probit model results.

This origin story of the probit model is well articulated in Gianola (1982). Note that Daniel Gianola was at University of Illinois when he wrote this paper.

What do we think about this?

The probit model (Bliss (1934a) and Bliss (1934b)) and threshold model origin story (Wright (1934) came before the celebrated Pitman-Koopman-Darmois theorem (proved independently by three different persons in 1935 and 1936).

The probit model predates exponential family GLMs (see Nelder and Wedderburn (1972)) and logistic regression.

One comment from a referee at a quantitative genetics journal: "There is nothing sacrosanct about the logistic link function."

Logistic Regression strikes back

The threshold model origin story of probit regression can be told through logistic regression.

Consider the logistic distribution with cdf

$$F(x; \mu, s) = \frac{1}{1 + \exp\left(-\frac{(x-\mu)}{s}\right)}$$

and set $\mu = 0$ and $s = 1$. With these specifications, the logistic pdf is

$$f(x; 0, 1) = \frac{\exp(-x)}{(1 + \exp(-x))^2}.$$

This distribution is symmetric and it forms a location-scale family.

What do we think about all of this?

Probit model fitting

The probit model has log likelihood

$$l(\beta) = \log \left\{ \prod_{i=1}^{n} \left[\Phi \left(\sum_j \beta_j x_{ij} \right) \right]^{y_i} \left[1 - \Phi \left(\sum_j \beta_j x_{ij} \right) \right]^{n_i - y_i} \right\}.$$

Differentiation with respect to β_j leads to,

$$\sum_i \left\{ \left[\frac{n_i \left[y_i - \Phi \left(\sum_j \beta_j x_{ij} \right) \right] x_{ij}}{\Phi \left(\sum_j \beta_j x_{ij} \right) \left[1 - \Phi \left(\sum_j \beta_j x_{ij} \right) \right]} \right] \phi \left(\sum_j \beta_j x_{ij} \right) \right\} = 0,$$

When the link function is not the canonical link there is no reduction of the data in the form of Fisher's notion of sufficieny.

The estimated asymptotic covariance matrix of $\hat{\beta}$ has the form

$$\widehat{\text{cov}}(\hat{\beta}) = \left(M^T \widehat{W} M\right)^{-1}.$$

For probit models, $\widehat{W}$ is the diagonal matrix with elements

$$w_i = \frac{n_i \left[\phi\left(\sum_j \hat{\beta}_j x_{ij}\right)\right]^2}{\left\{\Phi\left(\sum_j \hat{\beta}_j x_{ij}\right)\left[1 - \Phi\left(\hat{\beta}_j x_{ij}\right)\right]\right\}}.$$

Note: The Newton-Raphson algorithm yields the same MLE estimates but slightly different standard errors. For the information matrix inverted to obtain the asymptotic covariance matrix, Newton-Raphson uses observed information, whereas Fisher scoring uses expected information. These differ for link functions other than the canonical link.

Inverse propensity score weighting (IPW) in Causal Inference

One of the central goals of causal inference is to estimate the average treatment effect (ATE).

In its most simple presentation, we will assume that the treatment variable is binary.

The ATE measures the difference in mean outcomes between individuals assigned to the treatment and individuals assigned to the control.

However, there is a difficulty with estimation since individuals do not simultaneously receive the treatment and the control.

When the study is randomized this difficulty is mitigated because randomization ensures (when n is large enough) that baseline covariates which may influence the response have the same distribution in both the treatment and control groups.

In observational studies this may not be so, and the researcher may have no information about the treatment assignment mechanism.

One technique for estimating the ATE in causal studies is through *inverse propensity score weighting (IPW)*. The IPW technique makes use of a treatment assignment model. In the simple setting we will take a logistic regression model as our model for treatment assignment.

The idea of IPW is to create a pseudo population that exhibits balance in baseline covariates in an effort to mimic random assignment up to measured confounders.

Example

We will now go through an example. In this example we estimate the causal effect of online learning in STAT 200 at UIUC.

The response variable Y is the comprehensive 3-hour Final that was Scantron graded at the end of each semester, and the treatment variable A is the presence or absence of in-person lectures.

The online only course is considered the treatment and the regular in-person course with recorded lectures is the control.

We estimate the causal effect of online learning (or absence of in-person lecture) by estimating the ATE using inverse propensity score weighting methods. The form of this estimator of the ATE is

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left(w_i A_i Y_i - w_i(1-A_i)Y_i\right),$$

where the weights w_i are a functions of the propensity scores $\hat{p}_i = \widehat{P}(A_i = 1|X_i)$, $i = 1, \ldots, n$.

The propensity scores are each subjects conditional probability of belonging to the treatment group given their covariate information.

We will estimate the propensity scores using a logistic regression model.

We consider an alternative (the classical IPW) stable estimator of the ATE with weights of the form

$$w_i = \frac{\frac{A_i}{\hat{p}_i}}{\sum_{i=j}^{n} \frac{A_j}{\hat{p}_j}} + \frac{\frac{1-A_i}{1-\hat{p}_i}}{\sum_{j=1}^{n} \frac{1-A_j}{1-\hat{p}_j}}.$$

Let's load in the data and try it out.

```
dat <- read.csv("online.csv", header = TRUE)[, -1]
head(dat)
```

```
##   ObjExam Online ACTMath ACTMajor  ACT Gender International F17 S18 S19 Fa19 FR
## 1   86.75      1    31.0     31.5 30.0      0             0   1   0   0    0  0
## 2   91.57      1    35.6     31.5 33.2      0             0   1   0   0    0  0
## 3   86.75      1    35.6     33.5 33.9      0             1   1   0   0    0  0
## 4   96.39      0    36.0     33.5 35.0      0             0   1   0   0    0  0
## 5   78.31      0    35.0     31.5 30.0      0             1   1   0   0    0  1
## 6   67.47      1    33.0     30.5 33.0      1             0   1   0   0    0  0
##   SO JR
## 1  0  1
## 2  1  0
## 3  0  0
## 4  1  0
## 5  0  0
## 6  0  1
```

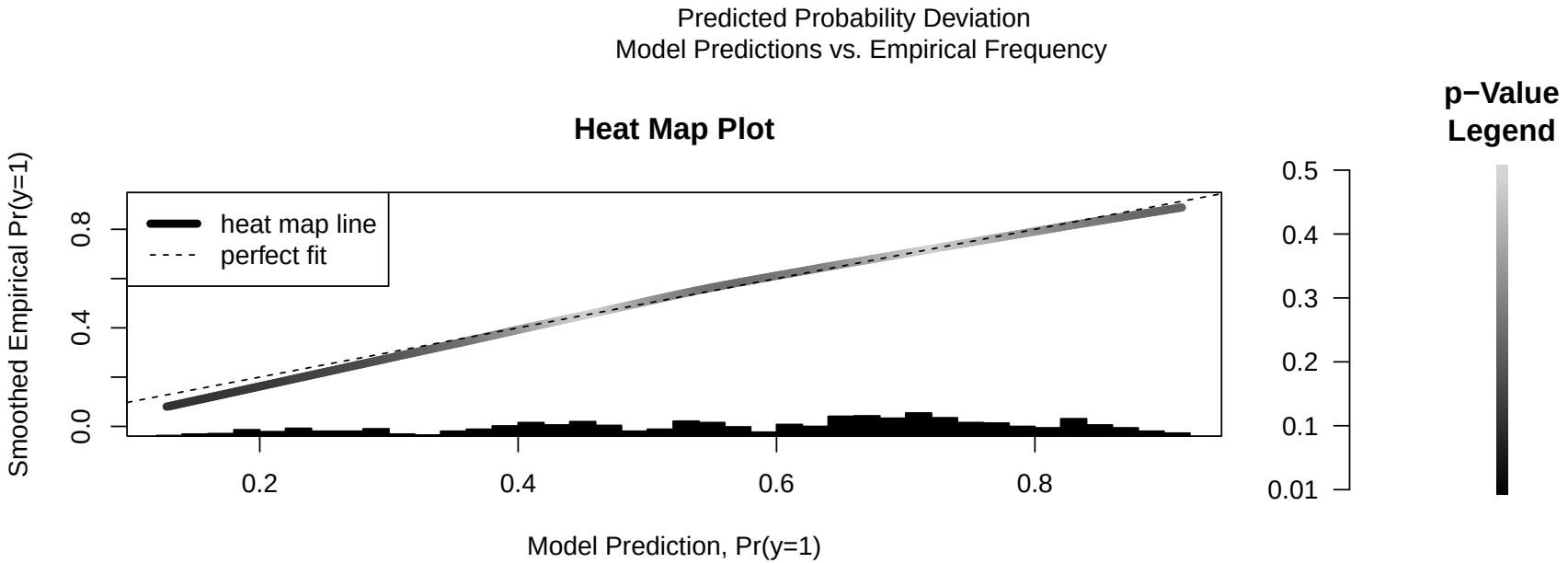
We specify the propensity score logistic regression model and obtain the estimated propensity scores.

```
dat_small <- dat %>%
  dplyr::select(Online, ACTMath, ACTMajor, ACT, Gender,
                International, F17, S18, S19, Fa19, FR, SO, JR)

# prop score model
m <- glm(Online ~., data = dat_small, family = "binomial")
preds <- predict(m, type = "response")
```

```
trt <- dat_small$Online
heatmap.fit(trt, preds)
```

```
##
## Calculating optimal loess bandwith...
## aicc Chosen Span =  0.9867739
##
## Generating Bootstrap Predictions...
##   |                                                                      |
```




```
##
##
## ****************************************
## 0% of Observations have one-tailed p-value <= 0.1
## Expected Maximum = 20%
## ****************************************
```

We now obtain the weights for the stabilized estimator and we check for the balance of weights in the online and in-person groups across the Gender and International variables

```
# estimate the stabilized IPW weights
weights_alt_trt <- 1 / sum(trt / preds) * trt /preds
weights_alt_notrt <- 1 / sum((1 - trt)/(1 - preds)) * (1-trt)/(1-preds)
dat <- data.frame(dat, weights = weights_alt_trt - weights_alt_notrt)

# check balance for gender and international (other variables are also balanced)
dat %>% group_by(Gender, Online) %>% summarise(sum(weights))
```

```
## # A tibble: 4 x 3
## # Groups:   Gender [2]
##   Gender Online `sum(weights)`
##    <int>  <int>          <dbl>
## 1      0      0         -0.563
## 2      0      1          0.574
## 3      1      0         -0.437
## 4      1      1          0.426
```

```
dat %>% group_by(International, Online) %>% summarise(sum(weights))
```

```
## # A tibble: 4 x 3
## # Groups:   International [2]
##   International Online `sum(weights)`
##           <int>  <int>          <dbl>
## 1             0      0         -0.587
## 2             0      1          0.596
## 3             1      0         -0.413
## 4             1      1          0.404
```

We now estimate the ATE corresponding to these weights.

```
ATE_alt <- sum(weights_alt_trt * dat$ObjExam) -
  sum(weights_alt_notrt * dat$ObjExam)
ATE_alt
```

```
## [1] 0.5775491
```

In the notes we also considered a double robust (DR) estimator estimator which may alleviate concerns with misspecification of the propensity score model.

```
# estimate DR version ATE
m_trt <- lm(ObjExam ~ ACTMath + ACTMajor + ACT + International + Gender +
            FR + SO + JR + F17 + S18 + S19,
            data = dat[trt == 1, ])
Y_trt <- predict(m_trt, newdata = dat)
m_notrt <- lm(ObjExam ~ ACTMath + ACTMajor + ACT + International + Gender +
              FR + SO + JR + F17 + S18 + S19,
              data = dat[trt == 0, ])
Y_notrt <- predict(m_notrt, newdata = dat)
ATE_DR <- mean( (dat$ObjExam * trt - (trt - preds) * Y_trt) / preds -
  (dat$ObjExam * (1 - trt) + (trt - preds)*Y_notrt) / (1 - preds))
ATE_DR
```

```
## [1] 0.4217033
```

Notes

The actual analysis of this data considered:

- multiple linear regression (no causal analysis)
- the stable IPW estimator considered here
- a double robust estimator considered here
- an outcome highly adaptive lasso (OHAL) approach that is robust to model misspecification

All of these approaches yielded similar results.

Detailed investigations of interactions and missing confounding variables were also considered.