PH 252D Spring 2018

Entrepreneurship in Uganda

Shruti Bathia, David Contreras Loya, William Krinsman 05/07/2018

Part I

Slides

slide 2 speaker notes: Notes for speaker: “Youth unemployment remains a serious policy challenge in many sub-Saharan African countries, including Uganda. In 2013, youth (aged 15 to 24) in sub-Saharan Africa were twice likely to be unemployed compared to any other age cohort. ” “A large population of Ugandans are underemployed i.e. being either highly skilled but working in low paying jobs or working part time. ”
Citation: https://www.brookings.edu/blog/africa-in-focus/2014/08/26/youth-unemployment-challenge-in-uganda-and-the-role-of-employment-policies-in-jobs-creation/
http://eprcug.org/blog/549-the-need-to-focus-on-the-growing-number-of-underemployed-persons-in-uganda-s-labour-force
slide 4 speaker notes: These studies examine the impact of training on existing entrepreneurs

slide 7 speaker notes: 200 secondary schools (1/3 of tota,, l number of secondary schools in Uganda) 8,080 students applied to the program and 7,421 complied with eligibility requirements.
Eligibility: (completeness of key baseline characteristics and no concurrent entrepreneurship or business training Power calculations showed that 1,200 students per arm were required, but sample size was incremented to account for attrition.
Then, 1,600 students were randomly assigned to hard skills, 1,600 to soft skills and 1,200 for the control group.
NOTE: 30% of those assigned to treatment didn’t go to the training at all, those who attended, attendance 94% of the days, so we’re reporting intent to treat.
There were no important differences in the observable characteristics of those (12%) who chose not to respond to the follow-up survey. Women were slightly likely more likely to respond to the follow-up.

slide 14: Is the knowledge and data sufficient , target parameter, Under the assumption that our observed data was generated by our SCM, O wil be a subset of endogenous variables X.

slide 20: Compliance was not perfect: Of those who attended, assistance averaged 94%

slide 21: Compliance was not perfect: Of those who attended, assistance averaged 94%

1 Background

Uganda, like most low-income countries, has a large share of youth who are either unemployed or underemployed. Living in economies where employment opportunities are scarce and self-employment is often the only option, youth need the right combination of human, financial, and social capital to improve their welfare. Younger people are often the largest demographic segment in these countries, which means that their well-being has especially important ramifications for the overall state of their countries’ economies.

Many governments recognize that their economy would benefit from better-trained entrepreneurs. Uganda and 22 other African countries have mainstreamed entrepreneurship training in high school through support from the International Labor Organization (ILO). Other countries are developing short training programs in entrepreneurship, while yet other countries are expanding university level entrepreneurship training. However, the curricula in all of these programs are based primarily on hard skills and ignore the potential contributions of soft skills to improved economic outcomes.

This proposed research seeks to address a gap in development literature by focussing on which specific business training techniques work. There have been a number of experimental business training evaluation studies including Karlan and Valdivia (2011)[4] and Valdivia (2011)[5] in Peru, Drexler et al. (2014)[3] in the Dominican Republic, Berge et al. (2011)[1] in Tanzania. These studies confirm that business training leads to improvements in knowledge of good business practices. However, these studies examine the impact of training on existing entrepreneurs. In Sri Lanka de Mel, McKenzie, and Woodruff (2012)[2] examine the effects of an ILO business training program on business success of both existing female entrepreneurs and the general population of women. The proposed project wishes to expand on this research.

More specifically, we want to investigate if entrepreneurial training affects labor market outcomes by a) inducing individuals to start businesses sooner after graduation of secondary school and b) increasing revenues and profits for those businesses. We measure business creation and financial performance in a sample of 3,893 Ugandans between 22-30 years old who were eligible to receive a three-week, post-secondary intensive training camp on entrepreneurship skills. We will study economic outcomes of individuals under a non-parametric framework to estimate their treatment-specific counterfactual outcomes. We hope to answer the following questions:

- Does entrepreneurial training (of any kind) increase the likelihood of starting a business after graduation from high school?
- Does entrepreneurial training (of any kind) increase business monthly revenue?
- Does entrepreneurial training (of any kind) increase business monthly profit?

All of the economic outcomes listed above are proxies for the success of entrepreneurial training in improving the welfare of young persons who might otherwise be unemployed or underemployed. For the purposes of this initial analysis, we will pool both treatment arms (hard-skills and soft-skills) into a single group, thereby ignoring any potential differences between the two types of training..

2 Experimental Design

We interviewed 4,400 individuals at baseline, and we reached 3,891 during the follow-up study 4 years after. Our baseline covariates W_0 include basic sociodemographic characteristics such as age, gender, region of residence, and household socio-economic level; several measures of cognitive development, e.g. Raven score; personality constructs (Big 5); and time and risk preferences. Distance from home village to training site was also recorded for all individuals. We observe treatment status A labeled as $A = 0$ for no treatment and $A = 1$ for treatment. At follow-up, we obtained information about every economic activity undertaken in the period after graduation from high school and time of the follow-up interview (April 2016). Our outcomes Y are (1) a binary indicator for whether the individual started a business, (2) the logarithm of monthly revenue measured in USD, and (3) the logarithm of monthly profit measured in USD. Note that outcomes (2) and (3) only apply to those individuals who actually started a business.

The target population was youth in Uganda who graduated high school and are in the job market. The sample consisted of students enrolled in the last year of high school in 4 regions of Uganda in 2013. Approximately, 40% of the sample attended schools in the West, 20% in Jinja, 20% in Mbale, and 20% in the North. The study was designed to be nationally representative with both students and teachers assigned to one of three groups (hard skills, soft skills, control) randomly. Students were recruited from 200 secondary schools, which represents a third of the total number of full time secondary schools in Uganda. Students interested in the program were asked to fill out an application form and a baseline survey. In total $8,080$ students applied to the program and of those $7,431$ complied with eligibility requirements (completeness of key baseline characteristics and no concurrent entrepreneurship or business training).

Power calculations showed that $1,200$ students per arm were required, but sample size was incremented to account for attrition. We drew a random sample of $4,400$ students out of the eligible pool of $7,421$. Then, $1,600$ students were randomly assigned to hard skills training, $1,600$ students were randomly assigned to soft skills, and $1,200$ students were randomly assigned to the control group. At each step of the sampling process we stratified by both school and gender to avoid confounding and ensure a well-balanced design.

A two-arms intervention was implemented: a 3-week intensive entrepreneurship camp with a strong emphasis on (1) soft skills and (2) hard skills. All students had a basic overview entrepreneurship and worked on a business plan during the 3-week course. The intervention was implemented in May 2013.

Students in the hard skills program focused on financial decision making, while the soft skills arm focused on abilities such as negotiation and communication. The curricula for the training was designed by the International Labor Organization and the Haas Business School.

Teachers were recruited, hired and trained by Educate! a non-profit organization. Teachers were randomly assigned to training site, school and classroom. Each of the 20 host schools was staffed with 3 teachers: 2 regular curriculum instructors (hard or soft skills), who both taught the regular curriculum, and 1 instructor who taught the business plan curriculum exclusively. Assignment was stratified by language and ability. The principal investigators of this study are Paul Gertler and Dana Carney at UC Berkeley.

Treatment was assigned randomly, i.e. using a random number generator. This was for identifiability of results.

Overall about one-third of the study participants are female. On average, those taking part in the study are 20 years old.

The sample was balanced across all 3 arms of the study (no treatment, soft-skills treatment, and hard skills treatment). 9 of 144 p-values were less than 0.10. The characteristics of the teachers were balanced as well. Of everyone assigned to treatment (hard- or soft-skills), 67.4% participated in the training. None of the controls participated in the training. Our sample consists of $1,021$ controls, $1,448$ individuals assigned to *hard* skills, and $1,422$ individuals assigned to *soft* skills. Roughly $2/3$ of the sample started a business during the recall period, and average monthly revenues and profits were 957 and 501 USD (adjusted for purchasing power parity, PPP).

3 Limitations

Even though assignment to treatment was randomized, compliance with treatment was not perfect (i.e. not every individual assigned to treatment attended the training). Moreover, we were able to reach approximately c88% of the original sample in the follow-up interview. Therefore, estimation of causal effects in this setting entails dealing with a potential selection problem, because individuals who did not attend the training, or individuals who were lost to follow-up could differ in observable and unobservable characteristics correlated with the outcomes. Fortunately, we have baseline covariates of those were lost to follow-up for the original $4,400$ individuals. By fully utilizing all the available baseline covariates, our aim is to estimate a double robust locally efficient substitution estimator that will be consistent and asymptotically linear if the selection mechanism is consistently estimated or if we can treat assignment to treatment and attrition as independent events (i.e. no differential attrition between treatment and control).

An empirical strategy to deal with this censoring issue is to model assignment to treatment A and attrition Δ as a single intervention node by estimating its joint distribution $f_{A,\Delta}(A, \Delta)$.

4 Causal Analysis

For each of the three outcomes, the target causal parameter is the Average Treatment Effect, which is the difference in the expected counterfactual if all recruited students had taken the entrepreneurial training and the expected counterfactual if none of the students were assigned to the treatment.

$$\Psi^{\mathcal{F}}(\mathbb{P}_{U,X}) = \mathbb{E}_{U,X}(Y_1) - \mathbb{E}_{U,X}(Y_0)$$

Because of the design of the experiment as an RCT (randomized control trial), we can conclude that A is a function of U_A only, so that there must be an exclusion restriction between W and A. See p. 24 of [6] for confirmation of this claim. A second consequence of our randomizing the intervention node A is that we may assume that U_A is independent of U_Y and of U_W. We can test this assumption for U_A and U_W

by using a balance table. There is no way for us to test the independence assumption of U_A and U_Y. This corresponds to believing that there are no unmeasured factors that predict both A and the outcome Y, which is often called the no unmeasured confounders assumption. This independence assumption also means that there is no backdoor path from Y to A. Therefore our causal estimand is identifiable from the statistical estimand. Our structural causal assumes that the observed data were generated by the following actions:

1. Drawing unobservable $U = (U_A, U_Y, U_W)$ from some probability distribution $\mathbb{P}_U$ ensuring that U_A is independent of U_Y, given W.
2. Generating W as a deterministic function of U_W.
3. Generating A as a deterministic function of W and U_A.
4. Generating Y as a deterministic function of W, A, and U_Y.

The time ordering of the variables is $W \rightarrow A \rightarrow Y$, while for the causal ordering, both W and A precede Y, and neither W nor A precede each other. This model is illustrated in the figure 1. Note that there are no assumptions on functional forms in our model. Nevertheless, because of the randomization of U_A via a random number generator with a specified distribution, our model is technically speaking only semi-parametric, and not completely non-parametric.

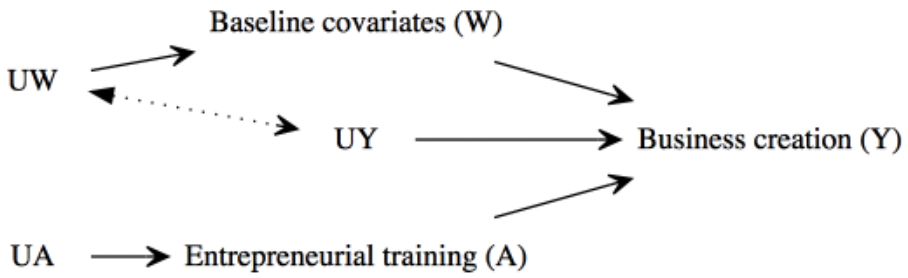

Figure 1: Structural causal model

5 Variables of interest

Variable type	Variable name	Description
Y	Business creation	=1 if respondent started a business after graduation from high school
Y	Log of revenue	Monthly revenue from all self-employment activities
Y	Log of profit	Monthly profit from all self-employment activities
A	Treatment	=1 if participated in entrepreneurial training
W	Sociodemographic characteristics	Gender, age, parent's income source and education level, boarding student, perceived socioeconomic level
W	Cognitive skills	Raven score, math score, GPA, O-level score, previous exposure to entrepreneurship
W	Risk and time preferences	Present-bias and time-inconsistency scores
W	Personality characteristics	Big 5 (extroversion, emotional stability, openness, conscientiousness, agreeableness), leadership, perceived control, anxiety, pro-social behavior, and more.

6 Interpretation

7 Future Directions

In future work we would like to investigate the differences in effectiveness, if any, between the hard-skills and soft-skills training. For the purposes of this initial analysis, we pooled both treatment pools into a single group. However, understanding the differences between the two treatment modes would have real policy implications: most existing entrepreneurship training programs only employ hard-skills training. Governments interested in effectively training their entrepreneurs would like to know whether hard-skills training, soft-skills training, or a combination of both, is most likely to improve the economic outcomes of the trainees. This was a question we completely neglected in the present analysis.

Bibliography

[1] Lars Ivar Oppedal Berge, Kjetil Bjorvatn, and Bertil Tungodden. *Human and financial capital for microenterprise development: Evidence from a field and lab experiment*. Discussion paper. Norges Handelshøyskole (Norwegian School of Economics and Business Administration), Institutt for Samfunnsøkonomi (Department of Economics), Jan. 21, 2011. URL: http://www.sv.uio.no/esop/english/publications/unpublished-works/working-papers/2011/tungodden%202011%20tanzania.pdf.

[2] Suresh De Mel, David McKenzie, and Christopher Woodruff. *The demand for, and consequences of, formalization among informal firms in Sri Lanka*. Working Paper 18019. National Bureau of Economic Research, Apr. 2012. DOI: 10.3386/w18019. URL: http://www.nber.org/papers/w18019.

[3] Alejandro Drexler, Greg Fischer, and Antoinette Schoar. "Keeping It Simple: Financial Literacy and Rules of Thumb". In: *American Economic Journal: Applied Economics* 6.2 (Apr. 2014), pp. 1–31. DOI: 10.1257/app.6.2.1. URL: http://www.aeaweb.org/articles?id=10.1257/app.6.2.1.

[4] Dean Karlan and Martín Valdivia. "Teaching Entrepreneurship: Impact of Business Training on Microfinance Clients and Institutions". In: *Review of Economics and Statistics* 93.2 (May 2011), pp. 510–527.

[5] Martín Valdivia. *Traning or Technical Assistance? A Field Experiment to Learn what Works to Increase Managerial Capital for Female Microentrepreneurs*. Report. World Bank, Mar. 2011. URL: http://siteresources.worldbank.org/INTGENDER/Resources/336003-1303333954789/final_report_bustraining_BM_march31.pdf.

[6] Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media, 2011.

Chapter 1

Appendices

1 Code for Analysis

In [1]:
```r
# Note: for full reproducibility of results, we should have set the random seed earlier.
set.seed(518)
# The values generated are similar to those from the slides, but not the same.

rm(list=ls())
getwd()
options(scipen=10)

suppressMessages(   library(tmle))
suppressMessages(   library(ggplot2))
                    library(SuperLearner)
suppressMessages(   library(dplyr))
                    library(magrittr)
                   library(foreign)
                    library(ck37r)
suppressMessages(   library(sl3))
suppressMessages(   library(arm))
                    library(lattice)
                    library(caret)
suppressMessages(   library(data.table))
                    library(screening)
suppressMessages(   library(xgboost))
                    library(foreach)
suppressMessages(   library(glmnet))
```

In [2]:
```r
data <- read.dta("Data/SEED_endline_analysis.dta",
                 convert.factors=FALSE, convert.underscore=FALSE)
data <- data.frame(data)
```

In [3]:
```r
# List to hold the different column names.
(names=list(
  # Outcomes of interest
  outcome=c("ever_self_employed","log_tot"),

  # Treatment variable
  treatment="treated",

  # Adjustment covariates
```

```
  covars=c("treated","gender","age","q06_dayorboarding",
  "q25_family_business","q25a_wk_family_bus","timeprefs_patience",
  "riskbehavior","mathbusiness","leadership","perceivedcontrol","timeprefs_delta",
  "timeprefs_beta","prosocialbehavior","anxiety","selfconfidence",
  "big5extroversion","big5emostability","big5openness","big5conscientious",
  "big5agreeable","schoolacceptance","currfamwealthstep","tenyrwealthstep","takingriskstep",
  "ravenscore","father_educ2","father_educ3","father_educ4","father_educ5",
  "father_income2","father_income3","mother_income2","mother_income3",
  "type_house","q13_olevelscore2","q13_olevelscore34")
))
```

$outcome
'ever_self_employed' 'log_tot'

$treatment
'treated'

$covars
'treated' 'gender' 'age' 'q06_dayorboarding' 'q25_family_business' 'q25a_wk_family_bus' 'timeprefs_patience' 'riskbehavior' 'mathbusiness' 'leadership' 'perceivedcontrol' 'timeprefs_delta' 'timeprefs_beta' 'prosocialbehavior' 'anxiety' 'selfconfidence' 'big5extroversion' 'big5emostability' 'big5openness' 'big5conscientious' 'big5agreeable' 'schoolacceptance' 'currfamwealthstep' 'tenyrwealthstep' 'takingriskstep' 'ravenscore' 'father_educ2' 'father_educ3' 'father_educ4' 'father_educ5' 'father_income2' 'father_income3' 'mother_income2' 'mother_income3' 'type_house' 'q13_olevelscore2' 'q13_olevelscore34'

In [4]:
```
# Keep variables of interest

data <- subset(data, select=c(names$outcome, names$treatment, names$covars))
# Review missing values in id, outcome, treatment, and censoring variables.
# Outcome is the only variable that can have missing values.
colSums(is.na(data[, c(names$outcome, names$censoring, names$treatment)]))
```

ever_self_employed 0
log_tot 712
treated 0

In [5]:
```
# Dimensions of data set
dim(data)
```

3891 40

In [6]:
```
# Summary statistics of data set
summary(data)
```

```
 ever_self_employed    log_tot          treated         treated.1
 Min.   :0.0000     Min.   : 1.028   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000     1st Qu.: 6.580   1st Qu.:0.0000   1st Qu.:0.0000
 Median :1.0000     Median : 7.681   Median :1.0000   Median :1.0000
 Mean   :0.5474     Mean   : 7.593   Mean   :0.7376   Mean   :0.7376
 3rd Qu.:1.0000     3rd Qu.: 8.645   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000     Max.   :11.018   Max.   :1.0000   Max.   :1.0000
                    NA's   :712
     gender            age        q06_dayorboarding q25_family_business
```

```
Min.   :0.0000   Min.   :20.00   Min.   :0.0000    Min.   :0.0000
1st Qu.:0.0000   1st Qu.:22.00   1st Qu.:0.0000    1st Qu.:0.0000
Median :0.0000   Median :23.00   Median :1.0000    Median :1.0000
Mean   :0.3482   Mean   :23.51   Mean   :0.7396    Mean   :0.5193
3rd Qu.:1.0000   3rd Qu.:24.00   3rd Qu.:1.0000    3rd Qu.:1.0000
Max.   :1.0000   Max.   :38.00   Max.   :1.0000    Max.   :1.0000
                                 NA's   :147       NA's   :13
q25a_wk_family_bus timeprefs_patience  riskbehavior        mathbusiness
Min.   :0.0000     Min.   :0.0000     Min.   :-2.538293   Min.   :0.0000
1st Qu.:1.0000     1st Qu.:0.0000     1st Qu.:-0.692131   1st Qu.:0.5000
Median :1.0000     Median :0.0000     Median :-0.083936   Median :0.6667
Mean   :0.9276     Mean   :0.2765     Mean   :-0.002497   Mean   :0.5990
3rd Qu.:1.0000     3rd Qu.:0.3333     3rd Qu.: 0.656105   3rd Qu.:0.7500
Max.   :1.0000     Max.   :1.0000     Max.   : 2.965215   Max.   :1.0000
NA's   :1860
  leadership    perceivedcontrol timeprefs_delta     timeprefs_beta
Min.   :1.000   Min.   :1.000    Min.   :-3.299497   Min.   :-3.048114
1st Qu.:3.857   1st Qu.:4.167    1st Qu.:-0.662538   1st Qu.:-0.682101
Median :4.286   Median :4.333    Median : 0.001506   Median :-0.013883
Mean   :4.194   Mean   :4.337    Mean   : 0.001506   Mean   : 0.002516
3rd Qu.:4.571   3rd Qu.:4.667    3rd Qu.: 0.643275   3rd Qu.: 0.637755
Max.   :5.000   Max.   :5.000    Max.   : 3.363326   Max.   : 3.857732
NA's   :23      NA's   :14
prosocialbehavior    anxiety      selfconfidence  big5extroversion
Min.   :1.000     Min.   :1.000   Min.   :1.000   Min.   :1.000
1st Qu.:4.000     1st Qu.:1.889   1st Qu.:4.333   1st Qu.:2.000
Median :4.293     Median :2.333   Median :4.667   Median :3.000
Mean   :4.293     Mean   :2.391   Mean   :4.583   Mean   :2.733
3rd Qu.:4.714     3rd Qu.:2.875   3rd Qu.:5.000   3rd Qu.:3.500
Max.   :5.000     Max.   :5.000   Max.   :5.000   Max.   :5.000
                  NA's   :28      NA's   :37
big5emostability  big5openness   big5conscientious big5agreeable
Min.   :1.000    Min.   :1.000   Min.   :1.000     Min.   :1.00
1st Qu.:3.500    1st Qu.:3.500   1st Qu.:3.500     1st Qu.:3.00
Median :4.000    Median :4.151   Median :4.000     Median :3.50
Mean   :3.865    Mean   :4.151   Mean   :3.892     Mean   :3.62
3rd Qu.:4.500    3rd Qu.:5.000   3rd Qu.:4.500     3rd Qu.:4.00
Max.   :5.000    Max.   :5.000   Max.   :5.000     Max.   :5.00

schoolacceptance currfamwealthstep tenyrwealthstep  takingriskstep
Min.   :1.000    Min.   : 1.000    Min.   : 1.000   Min.   : 1.000
1st Qu.:4.000    1st Qu.: 4.000    1st Qu.: 7.000   1st Qu.: 5.000
Median :4.250    Median : 5.000    Median : 8.000   Median : 7.000
Mean   :4.268    Mean   : 4.776    Mean   : 8.015   Mean   : 6.756
3rd Qu.:4.750    3rd Qu.: 6.000    3rd Qu.: 9.000   3rd Qu.: 9.000
Max.   :5.000    Max.   :10.000    Max.   :10.000   Max.   :10.000
NA's   :91       NA's   :83        NA's   :81       NA's   :88
  ravenscore      father_educ2     father_educ3   father_educ4
Min.   : 0.000   Min.   :0.0000   Min.   :0.00   Min.   :0.0000
1st Qu.: 4.000   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:0.0000
Median : 6.000   Median :0.0000   Median :0.00   Median :0.0000
Mean   : 5.435   Mean   :0.1667   Mean   :0.13   Mean   :0.1838
3rd Qu.: 7.000   3rd Qu.:0.0000   3rd Qu.:0.00   3rd Qu.:0.0000
Max.   :10.000   Max.   :1.0000   Max.   :1.00   Max.   :1.0000
```

```
                NA's   :28       NA's   :28     NA's   :28
 father_educ5   father_income2   father_income3   mother_income2
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
Mean   :0.4072   Mean   :0.2924   Mean   :0.0384   Mean   :0.1553
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
NA's   :28       NA's   :37       NA's   :37       NA's   :15
mother_income3      type_house     q13_olevelscore2 q13_olevelscore34
Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
Median :0.00000   Median :1.0000   Median :0.0000   Median :0.0000
Mean   :0.03199   Mean   :0.8205   Mean   :0.4014   Mean   :0.4522
3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
NA's   :15        NA's   :24       NA's   :72       NA's   :72
```

In [7]:
```
# Remove observations missing their censoring time.
skip_vars <- c(names$treatment, names$outcome)
impute <- ck37r::impute_missing_values(data,
                                       skip_vars=skip_vars)
```

In [8]:
```
# Review missing data for all covariates.
# Only the outcome variable should have missing data at this point.
data <- impute$data

colSums(is.na(data))
```

ever_self_employed	0
log_tot	712
treated	0
treated.1	0
gender	0
age	0
q06_dayorboarding	0
q25_family_business	0
q25a_wk_family_bus	0
timeprefs_patience	0
riskbehavior	0
mathbusiness	0
leadership	0
perceivedcontrol	0
timeprefs_delta	0
timeprefs_beta	0
prosocialbehavior	0
anxiety	0
selfconfidence	0
big5extroversion	0
big5emostability	0
big5openness	0
big5conscientious	0
big5agreeable	0

schoolacceptance	0
currfamwealthstep	0
tenyrwealthstep	0
takingriskstep	0
ravenscore	0
father_educ2	0
father_educ3	0
father_educ4	0
father_educ5	0
father_income2	0
father_income3	0
mother_income2	0
mother_income3	0
type_house	0
q13_olevelscore2	0
q13_olevelscore34	0
miss_log_tot	0
miss_q06_dayorboarding	0
miss_q25_family_business	0
miss_q25a_wk_family_bus	0
miss_leadership	0
miss_perceivedcontrol	0
miss_anxiety	0
miss_selfconfidence	0
miss_schoolacceptance	0
miss_currfamwealthstep	0
miss_tenyrwealthstep	0
miss_takingriskstep	0
miss_father_educ2	0
miss_father_income2	0
miss_mother_income2	0
miss_type_house	0
miss_q13_olevelscore2	0

In [9]:
```
## Estimation of causal efects

Y1 <- data$ever_self_employed
Y2 <- data$log_tot[!is.na(data$log_tot)]

A1 <- data$treated
A2 <- data$treated[!is.na(data$log_tot)]

all_covars <- data[, colnames(data) %in% names$covars]

W <-  all_covars
W1 <- all_covars
W2 <- subset(data, !is.na(data$log_tot))
W2 <- W2[, colnames(data) %in% names$covars]

screen1 <-  screening(x=W1, y=Y1, method="holp", family="binomial", num.select=15)$screen
screen2 <-  screening(x=W2, y=Y2, method="holp", family="gaussian", num.select=15)$screen
screenA <-  screening(x=W,  y=A1, method="holp", family="binomial", num.select=15)$screen
screenA2 <- screening(x=W2, y=A2, method="holp", family="binomial", num.select=15)$screen
```

```
W1 <- W1[,screen1]
W2 <- W2[,screen2]

# William: moved this line here to make code work
screenA2 <- screening(x=W2, y=A2, method="holp", family="binomial", num.select=15)$screen
# screenA2 depends on W2, W2 was changed above, so old screenA2 can't be used to subset new W2

WA <-  W[ ,screenA]
WA2 <- W2[,screenA2]
```

In [10]:
```
# Fit glm model (base model, should have the worst performance)

logit2prob <- function(logit){
  odds <- exp(logit)
  prob <- odds / (1 + odds)
  return(prob)
}

model1 <- glm(formula=Y1 ~ A1, family="binomial")
summary(model1)
```

```
Call:
glm(formula = Y1 ~ A1, family = "binomial")

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.306  -1.306   1.054   1.054   1.224

Coefficients:
            Estimate Std. Error z value   Pr(>|z|)
(Intercept) -0.10784    0.06268  -1.720     0.0854 .
A1           0.40549    0.07317   5.542 0.00000003 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5359.0  on 3890  degrees of freedom
Residual deviance: 5328.2  on 3889  degrees of freedom
AIC: 5332.2

Number of Fisher Scoring iterations: 4
```

In [11]:
```
logit_control <- model1$coefficients[1]
logit_treated <- model1$coefficients[1] + 1*model1$coefficients[2]

(b1 <- logit2prob(logit_treated) - logit2prob(logit_control))
```

(Intercept): 0.10080197387954

In [12]:
```
model2 <- glm(formula=Y2 ~ A2, family="gaussian")
summary(model2)
```

```
Call:
glm(formula = Y2 ~ A2, family = "gaussian")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.6009  -1.0053   0.0917   1.0569   3.4584

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.49175    0.05073 147.679   <2e-16 ***
A2           0.13687    0.05895   2.322   0.0203 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 2.123154)

    Null deviance: 6756.7  on 3178  degrees of freedom
Residual deviance: 6745.3  on 3177  degrees of freedom
AIC: 11419

Number of Fisher Scoring iterations: 2
```

In [13]:
```
# Define our Super Learner library

g_library <- c("SL.mean",
               "SL.glm",
               "SL.glm.interaction")

Q_library <- c("SL.mean",
               "SL.glm",
               "SL.glm.interaction",
               #"SL.glmnet",
               #"SL.randomForest",
               #"SL.bartMachine",
               "SL.xgboost")
```

In [14]:
```
############################################################
# G-computation formula
############################################################

np_boot_gcomp <- function(Y, A, W, nrep, family){

  X <- cbind(A,W)
  print(colnames(X))
  # wrapped in suppressWarnings() to prevent excessive verbosity
  suppressWarnings(
    QbarSL <- SuperLearner(Y=as.numeric(Y),
                           X=X,
                           SL.library=Q_library,
                           family=family)
    )

  results <- rep(NA, nrep)
```

```
    n       <- NROW(Y)
    #stop("stop")
    for(i in 1:nrep){

      i_boot  <- sample(1:nrow(W), size=n, replace=TRUE)
      W_boot  <- X[i_boot,]
      W1_boot <- W0_boot <- W_boot

      W1_boot$A <- 1
      W0_boot$A <- 0

      #psi_bootstrap <- G_comp(Y = Y_b, A = A_b, W = W_b, family = family)
      # wrapped in suppressWarnings() to prevent excessive verbosity
      suppressWarnings(
        Qbar1W <- predict(QbarSL, newdata=W1_boot, type="response")$pred
        )

      # wrapped in suppressWarnings() to prevent excessive verbosity
      suppressWarnings(
        Qbar0W <- predict(QbarSL, newdata=W0_boot, type="response")$pred
        )

      psi_bootstrap <- (Qbar1W - Qbar0W)
      # wrapped in suppressWarnings() to prevent excess verbosity in output
      suppressWarnings(
      results[i]    <- psi_bootstrap
          )
    }
    return(results)
  }
```

In [15]:
```
# For business creation
g_comp_boot <- np_boot_gcomp(Y=Y1, A=A1, W=W1, nrep=100, family="binomial")
summary(g_comp_boot)

(b_iptw  <- mean(g_comp_boot))

(sd_iptw <- sd(  g_comp_boot))

t_stat <- b_iptw/sd_iptw
(p_val <- dt(t_stat, df=n-1, log=FALSE))

quantile(g_comp_boot, probs=c(0.025,0.975))
```

```
 [1] "A"                 "gender"            "age"
 [4] "q06_dayorboarding" "treated"           "type_house"
 [7] "mother_income2"    "father_income3"    "father_income2"
[10] "prosocialbehavior" "big5emostability"  "currfamwealthstep"
[13] "ravenscore"        "big5agreeable"     "leadership"
[16] "big5openness"


   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.07751 0.09683 0.10461 0.10326 0.11016 0.11403
```

0.103255428507584

0.00775579552103696

2.07434461855019e-23

2.5%	0.0873818847222731
97.5%	0.112134764063077

In [16]:
```
# For log of total earnings
tot_g_comp_boot <- np_boot_gcomp(Y=Y2, A=A2, W=W2, nrep=100, family="gaussian")
summary(tot_g_comp_boot)

(b_iptw  <- mean(tot_g_comp_boot))

(sd_iptw <- sd(  tot_g_comp_boot))

t_stat <- b_iptw/sd_iptw
(p_val <- dt(t_stat, df=n-1, log=FALSE))

quantile(tot_g_comp_boot, probs=c(0.025,0.975))
```

```
 [1] "A"                   "gender"              "q13_olevelscore34"
 [4] "big5emostability"    "timeprefs_delta"     "treated"
 [7] "q25_family_business" "q06_dayorboarding"   "q13_olevelscore2"
[10] "age"                 "tenyrwealthstep"     "timeprefs_beta"
[13] "leadership"          "father_educ4"        "mother_income3"
[16] "anxiety"
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.01737  0.09807  0.13171  0.13350  0.17054  0.24804
```

0.133499791152873

0.0541475573594245

0.0202236528538269

2.5%	0.0276082188190126
97.5%	0.230772835739166

In [17]:
```
############################################################
# IPTW
############################################################

iptw <- function(Y, A, X, family){

  n  <- NROW(Y)

  # wrapped in suppressWarnings() to prevent excessive verbosity
  suppressWarnings(
    propensity_score <- SuperLearner(Y=A,
                  X=X,
```

```
                        SL.library=g_library,
                        family=family)
    )

  # Obtain predicted probability of treatment
  # wrapped in suppressWarnings() to prevent excessive verbosity
  suppressWarnings(
    pred_g1W <- predict(propensity_score, newX=X, type='response')$pred
    )

  # Probability of not being treated
  pred_g0W <- 1 - pred_g1W

  # Create vector gAW
  gAW <- rep(NA, n)
  gAW[A==1] <- pred_g1W[A==1]
  gAW[A==0] <- pred_g0W[A==0]

  # Create vector with inverse of predicted probability
  wt <- 1/gAW

  # Implement stabilized IPTW estimator (a.k.a. the modified Horvitz-Thompson estimator)
  Psi_hat <- mean(as.numeric(A==1)*wt*Y)/mean(as.numeric(A==1)*wt) -
             mean(as.numeric(A==0)*wt*Y)/mean(as.numeric(A==0)*wt)
  return(Psi_hat)
  }

np_boot <- function(Y, A, X, family, nrep){

  results <- rep(NA, nrep)
  n       <- NROW(Y)
  df      <- cbind(Y,A,X)

  for(i in 1:nrep){

    i_boot       <- sample(1:nrow(df), size=n, replace=TRUE)
    df_bootstrap <- df[i_boot,]

    Y_b <- df_bootstrap[,1]
    A_b <- df_bootstrap[,2]
    W_b <- subset(df_bootstrap, select=-c(1,2))

    psi_bootstrap <- iptw(Y=Y_b, A=A_b, X=W_b, family=family)
    # added call to suppressWarnings() to avoid excess verbosity
    suppressWarnings(
    results[i]    <- psi_bootstrap
        )

  }
  return(results)
}
```

In [18]:
```
# IPTW for business creation
(ate_iptw <- iptw(Y=Y1, A=A1, X=WA, family="binomial"))
```

```
        # added argument 'family = "binomial"'
        # to avoid error 'argument "family" is missing, with no default'
        iptw_bootstrap <- np_boot(Y=Y1, A=A1, X=WA, nrep=100, family="binomial")
        summary(iptw_bootstrap)

        (b_iptw  <- mean(iptw_bootstrap))

        (sd_iptw <- sd(  iptw_bootstrap))

        t_stat <- b_iptw/sd_iptw
        (p_val <- dt(t_stat, df=n-1, log=FALSE))

        quantile(iptw_bootstrap, probs=c(0.025,0.975))
```

0.10080197387954

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.05579 0.09050 0.10286 0.10203 0.11563 0.14312
```

0.102033773035489

0.018674796478849

0.000000753327084341801

2.5%	0.0684991007238237
97.5%	0.136492148157286

In [19]:
```
# IPTW log total earnings

        (total_earn_iptw <- iptw(Y=Y2, A=A2, X=WA2, family="gaussian"))

        total_iptw_bootstrap <- np_boot(Y=Y2, A=A2, X=WA2, nrep=100, family="gaussian")
        summary(total_iptw_bootstrap)

        (b_iptw  <- mean(total_iptw_bootstrap))

        (sd_iptw <- sd(  total_iptw_bootstrap))

        t_stat <- b_iptw/sd_iptw
        (p_val <- dt(t_stat, df=n-1, log=FALSE))

        quantile(total_iptw_bootstrap, probs=c(0.025,0.975))
```

0.136872104521482

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.02517  0.09730  0.13004  0.12764  0.16301  0.22453
```

0.127635350433552

0.0523966603030267

0.02166695638499

2.5%	0.0269823320969634
97.5%	0.219980771351561

In [20]:
```
#####################################################
# TMLE
#####################################################

# Business creation
(tmle <- tmle(Y=as.numeric(Y1),
                  A=as.numeric(A1),
                  W=W1,
                  gform="A~1",
                  family="binomial",
                  #g.SL.library = g_library,
                  Q.SL.library=Q_library,
                  fluctuation="logistic") #,
                  #V=10)
)
```

```
Additive Effect
  Parameter Estimate:  0.10723
  Estimated Variance:  0.00029505
             p-value:  0.00000000043039
   95% Conf Interval: (0.073562, 0.1409)

Additive Effect among the Treated
  Parameter Estimate:  0.10723
  Estimated Variance:  0.0002948
             p-value:  0.00000000042335
   95% Conf Interval: (0.073576, 0.14088)

Additive Effect among the Controls
  Parameter Estimate:  0.10754
  Estimated Variance:  0.00029582
             p-value:  0.0000000004043
   95% Conf Interval: (0.073826, 0.14125)

Relative Risk
  Parameter Estimate:  1.2289
             p-value:  0.0000000040952
   95% Conf Interval: (1.1473, 1.3164)

              log(RR):  0.20615
   variance(log(RR)):  0.0012291

Odds Ratio
  Parameter Estimate:  1.5394
             p-value:  0.00000000048845
   95% Conf Interval: (1.3438, 1.7635)

              log(OR):  0.43142
```

```
    variance(log(OR)):  0.0048065
```

```
In [21]: # Log of total earnings
         (tot_tmle <- tmle(Y=as.numeric(Y2),
                           A=as.numeric(A2),
                           W=W2,
                           gform="A~1",
                           family="gaussian",
                           #g.SL.library = g_library,
                           Q.SL.library=Q_library,
                           fluctuation="logistic") #,
                           #V=10)
         )
```

```
 Additive Effect
   Parameter Estimate:  0.14313
   Estimated Variance:  0.0030896
              p-value:  0.010024
    95% Conf Interval: (0.034184, 0.25207)

 Additive Effect among the Treated
   Parameter Estimate:  0.14313
   Estimated Variance:  0.0030835
              p-value:  0.009951
    95% Conf Interval: (0.034291, 0.25197)

 Additive Effect among the Controls
   Parameter Estimate:  0.14313
   Estimated Variance:  0.0031101
              p-value:  0.010273
    95% Conf Interval: (0.033823, 0.25243)
```