8 Voter Persuasion

Business Understanding and Motivation

Einleitung

„Every method we have discussed so far is able to detect patterns in past data. […] However, when you are analyzing business and economic systems … **you need to be able to predict a future that will be different from the past because you are going to take actions that makes it different**. A decision that changes your product, marketing, or pricing **creates a new data generating process** that can break the … patterns that you‘ve seen in the past“

(Taddy, 2019, 127).

Hence, we are in a setting that “requires counterfactual prediction”. Predict something based on a DGP which is not yet in place, is not yet a fact.

“The gold standard for measuring the effect of an action is an experiment”. It is also the gold standard for “**estimating counterfactuals**” (Taddy, 2019, 128).

Randomized Control Trials are abbreviated as RCT and often called **A/B experiments**. For example, you randomize your website visitors into groups `A'

(control) and `B' (treatment). Those in A see the current website, while those in B see a new layout. If you are good at doing this the treatment effect in general can be estimated as the difference in means of the response

(Taddy, 2019, 128).

Einleitung

Könnte man alle Einflussfaktoren auf A und B messen, dann würde die Aktion (Treatment) eine weitere (Dummy-) Variable darstellen und wir könnten als ersten Ansatz eine multiple lineare Regression wagen (-> Bsp. Regression „wages“ aus Bachelor/Verbeek). Jedoch kennen wir nicht alle Einflussfaktoren auf A und B und können sie oft auch nicht messen.

Deshalb wird in **Experimentaldaten** durch die **Zufallsaufteilung** (Randomization) auf Experimental (B) - und Kontrollgruppe (A) versucht, mögliche Drittvariableneinfüsse zu mitigieren.

„Von einem **experimentellen Versuchsdesign** sprechen wir, wenn 3 Bedingungen vorliegen:

1. Es werden … zwei experimentelle Gruppen gebildet.
2. Die Versuchspersonen werden den experimentellen Gruppen nach einem Zufallsverfahren zugewiesen (Randomisierung).
3. Die unabhängige Variable wird vom Forscher ‚manipuliert‘“ Diekmann (2014*, 337).

Sollten Sie ein RCT durchführen wollen, lessen Sie unbedingt Diekmann (2014, Kap. VIII ff) und Taddy (2019, Ch. 5&6). Auch sollten Sie die verschiedenen Arten der Wahrscheinlichkeitsauswahl kennen (Diekmann, 2014, 380ff).

*Diekmann, A., 2014, Empirische Sozialforschung.

Business Understanding

Einleitung

Wir verwenden einen Datensatz aus Shmueli (2020, 335). Es handelt sich um ein Randomized Control Trials (A/B experiment). 10000 Wähler aus dem U.S.-Bundesstaat Delaware sind zufällig ausgewählt worden und die ersten 5000 im Datensatz haben eine Botschaft (einen Brief = Treatment) erhalten, die sie zur Wahl der Demokraten veranlassen soll (konkret des demokratischen Kandidaten Smith).

Die Daten sind zudem bereits aufbereitet worden:
Das Vorhandensein des Treatment signalisiert die Variable **MESSAGE_A** im Datensatz.

Einige Zeit nach dem Erhalt der Botschaft wird geprüft, ob sich der Wähler Richtung Demokraten bewegt hat. We conduct “a post-message survey … to measure whether each voter’s opinion … has shifted in a positive direction. A binary variable MOVED_A … [indicates] **whether opinion has moved in a Democratic direction** (1) or not (0)” Shmueli (2020, 336).

In diesem Fall ist deshalb **MOVED_A** = 1, sonst 0.

MM1 Hier nur Business Understanding gewünscht?
Moser, Moritz; 2024-02-08T10:53:36.148

Business Understanding and Motivation

Business Analytics kann als angewandte Data Science in der Domäne Betriebswirtschaft verstanden werden.

Business Analytics ist die **Kompetenz, verfügbare unternehmensinterne und –externe Daten so zu analysieren, dass konkrete betriebswirtschaftliche Frage/- Problemstellungen evidenzbasiert gelöst werden können**, um für Unternehmen (dauerhafte) Wettbewerbsvorteile generieren zu können.

Wie also lautet die Fragestellung?

Forschungsfrage

Primär:

Hat sich das Versenden der Botschaften „PRO Demokraten“ gelohnt?

Sekundär:

What is the uplift for each voter, i.e. increase in propensity to vote Democrats?

Quelle in Anlehnung an Seiter, M. Business Analytics: Effektive Nutzung fortschrittlicher Algorithmen in der Unternehmenssteuerung (German Edition) (S.1). 2017, Vahlen. Kindle-Version.

Data Understanding & Preparation

Beschreibung

Datendeskription folgt Shmueli (2020, 335):

Regressand **MOVED_A** und Treatment Dummy **MESSAGE_A** sind bekannt.

Voter_ID benötigen wir für die praktische Berechnung des individuellen Uplifts.

AGE zeigt, dass auch Hundertjährige dabei sind.

NH_WHITE ist Prozentsatz der Weissen im Haushalt. D.h. im Extrem 100% = rein weisser Haushalt, jedoch in Daten maximal 99%. Würde man bei Datenvollzugriff recherchieren.

COMM_PT = Prozentsatz der arbeitenden Bevölkerung im Stadtteil, der ÖPNV nutzt. Maximum von 19% kein Wunder! Delaware ist ein ländlicher Bundesstaat an der Ostküste der Vereinigten Staaten von Amerika. Es gibt keine grossen Städte.

H_F1 = Single Female Household sind 11,6% der Fälle.

```
summary(dfrm_KUNDE)

##     MOVED_A          VOTER_ID           AGE            NH_WHITE
##  Min.   :0.0000   Min.   :    13   Min.   : 18.00   Min.   :23.00
##  1st Qu.:0.0000   1st Qu.:152719   1st Qu.: 36.00   1st Qu.:58.00
##  Median :0.0000   Median :310684   Median : 51.00   Median :65.00
##  Mean   :0.3734   Mean   :313104   Mean   : 50.87   Mean   :66.37
##  3rd Qu.:1.0000   3rd Qu.:470629   3rd Qu.: 64.00   3rd Qu.:85.00
##  Max.   :1.0000   Max.   :636334   Max.   :100.00   Max.   :99.00
##     COMM_PT            H_F1          REG_DAYS         PR_PELIG
##  Min.   : 0.000   Min.   :0.0000   Min.   :    0.0   Min.   :  0.00
##  1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.:  893.8   1st Qu.:  0.00
##  Median : 2.000   Median :0.0000   Median : 3353.5   Median :  0.00
##  Mean   : 3.873   Mean   :0.1163   Mean   : 4172.8   Mean   : 14.89
##  3rd Qu.: 4.000   3rd Qu.:0.0000   3rd Qu.: 6423.5   3rd Qu.: 33.00
##  Max.   :19.000   Max.   :1.0000   Max.   :21187.0   Max.   :100.00
##     E_PELIG          POLITICALC       MESSAGE_A
##  Min.   :  0.00   Min.   :0.0000   Min.   :0.0
##  1st Qu.: 20.00   1st Qu.:0.0000   1st Qu.:0.0
##  Median : 40.00   Median :1.0000   Median :0.5
##  Mean   : 40.54   Mean   :0.5311   Mean   :0.5
##  3rd Qu.: 60.00   3rd Qu.:1.0000   3rd Qu.:1.0
##  Max.   :100.00   Max.   :7.0000   Max.   :1.0
```

REG_DAYS = „Days since voter registered at current address“

PR_PELIG = „Voted in x % of non-presidential primaries“

E_PELIG = „Voted in x % of any primaries“

POLITICALC = „Is there a political contributor in the home? Yes = 1“. Erstaunlich hoher Anteil.
Maximalwert von 7 sieht nicht nach Dummy aus. Deshalb ignorieren wir diese Grösse!

Modeling

Einleitung

Aus den Grundkursen in wissenschaftlichen Methoden kennen Sie den Test für den Vergleich zweier Mittelwerte. Der Anteil von „Bewegten“ in der Experimentalgruppe A ist 34,4% und in der Kontrollgruppe B 40,2%.

```
percentageGroupB = sum(dfrm_KUNDE[1:5000,1])/5000
percentageGroupA = sum(dfrm_KUNDE[5001:10000,1])/5000
percentageGroupA

## [1] 0.3444

percentageGroupB

## [1] 0.4024
```

Die 6% Unterschied sehen gross aus. Mit einem t-Test bei 10000-2 Freiheitsgraden können wir prüfen, ob diese Differenz zweier Mittelwerte signifikant von Null verschieden ist. Bei unabhängigen Stichproben wird auch zwischen unbekannten, aber gleichen Varianzen und unbekannten ungleichen Varianzen unterschieden. Man könnte natürlich vorneweg die Gleichheit der Varianzen mit einem F-Test prüfen, aber wir wollen diesen simplen Ansatz nicht zu sehr vertiefen. Deshalb unterstellen wir unbekannte und ungleiche Varianzen.

Einleitung

Für die Nullhypothese H_0: $\mu_1 = \mu_2$, gilt die t-Teststatistik:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

T ist unter H_0 approximativ
t_f-verteilt

In R:

```
#test it via t test
ATE = percentageGroupB - percentageGroupA #ATE = average treatment effect
ATE_stdev =
sqrt(1/5000*var(dfrm_KUNDE[1:5000,1])+1/5000*var(dfrm_KUNDE[5001:10000,1]))
#formula 5.3 taddy & Page 537 Rinne Taschenbuch der Statistik
T = ATE / ATE_stdev
T #degfree = 10000-2 # hence, may use normal dist

## [1] 6.005569
```

Die kritischen Grenzen für t-Werte kennen wir schon vom Test der H0: ß=0.

Interpretieren Sie! Lohnt sich das Treatment?

Antwort auf primäre Forschungsfrage

Nicht nur die Nullhypothese H_0: $\mu_1 = \mu_2$ können wir komfortabel mit einer R Funktion testen:

```
t.test(dfrm_KUNDE[1:5000,1], dfrm_KUNDE[5001:10000,1], alternative =
'two.sided', mu = 0)

##
##  Welch Two Sample t-test
##
## data:  dfrm_KUNDE[1:5000, 1] and dfrm_KUNDE[5001:10000, 1]
## t = 6.0056, df = 9988.1, p-value = 1.973e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.03906896 0.07693104
## sample estimates:
## mean of x mean of y
##    0.4024    0.3444
```

Liegt die Null im 95% Konfidenzintervall?

Welche Teststatistik wird ausgewiesen?

Welcher p-value?

Interpretieren Sie! Lohnt sich das Treatment?

Modeling

Uplift

Wir wollen zukünftig die Wähler adressieren, bei denen der Anstieg in Wahrscheinlichkeit nach Erhalt der Botschaft (=Uplift) am grössten ist!

Übliches Vorgehen nach Shmueli (2020, 337):

- "Fit a classification model (0/1 – respond or not), with all predictor variables, including which message was sent
- Run the model for all voters twice
  1. With original data
  2. With message predictor reversed
- You now have two propensity scores for each voter
  - One as if they got message A
  - One as if they got message B
- **Propensity for favorable response with B minus propensity with A is the uplift for B over A**

Note: Also used in marketing to "microtarget" different marketing messages appropriately"

Modeling

Political Awareness

Wir wollen zukünftig die Wähler adressieren, bei denen der Anstieg in Wahrscheinlichkeit nach Erhalt der Botschaft (=Uplift) am grössten ist!

Wir wollen aber nicht nach Alter, Geschlecht oder Hautfarbe diskriminieren. Auch weil wir erahnen, dass dies innerhalb der Partei problematisch werden könnte.
Deswegen beschränken wir uns auf folgende Regressoren in einer Logit Regression:

MESSAGE_A
COMM_PT
REG_DAYS
PR_PELIG
E_PELIG

Wie beurteilen Sie solche Selbstbeschränkungen in der Wissenschaft?

Hier geht es jedoch um einen praktischen Anwendungsfall.

Modeling

Logit

Wir erhalten:

```
base_accu = giveAccuracy(d,d_hat)
base_accu #higher than benchmark of 50%
```

```
## [1] 65.07
```

```
logit_base = glm(MOVED_A ~ MESSAGE_A + COMM_PT + REG_DAYS + PR_PELIG +
E_PELIG,
                 family = "binomial", data = dfrm_KUNDE)
summary(logit_base)
```

```
##
## Call:
## glm(formula = MOVED_A ~ MESSAGE_A + COMM_PT + REG_DAYS + PR_PELIG +
##     E_PELIG, family = "binomial", data = dfrm_KUNDE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7445  -0.9594  -0.8207   1.3156   1.7673
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.238e+00  5.399e-02 -22.923  < 2e-16 ***
## MESSAGE_A    2.663e-01  4.224e-02   6.303 2.91e-10 ***
## COMM_PT      6.688e-02  4.217e-03  15.860  < 2e-16 ***
## REG_DAYS    -1.130e-05  5.304e-06  -2.130  0.03318 *
## PR_PELIG    -2.702e-03  1.003e-03  -2.695  0.00705 **
## E_PELIG      9.817e-03  1.022e-03   9.607  < 2e-16 ***
```

```
exp(coef(logit_base))
```

```
## (Intercept)   MESSAGE_A     COMM_PT    REG_DAYS    PR_PELIG     E_PELIG
##   0.2900698   1.3050972   1.0691663   0.9999887   0.9973017   1.0098656
```

Interpretieren Sie!

Note: If we control for other factors, impact of MESSAGE is even higher! -> Wages

Modeling

Uplifts

Wir senden nun – hypothetisch – nur der anderen Gruppe eine Botschaft und nutzen das geschätzte Modell, um den Unterschied zu messen.

```
head(d_hat)

##         1         2         3         4         5         6
## 0.3057288 0.2875339 0.5295597 0.4047943 0.4108482 0.4445762

head(d_hat_reversed)

##         1         2         3         4         5         6
## 0.2522889 0.2361924 0.4630920 0.3425828 0.3482506 0.3801557
```

Beim ersten Wähler würde das Weglassen der Botschaft dazu führen, das das Logit Modell nur noch 25% WS vorhersagt. Der Uplift beträgt also grob 5%.

```
uplift = matrix(d_hat - d_hat_reversed, 10000, 1)
head(uplift)

##             [,1]
## [1,] 0.05343989
## [2,] 0.05134148
## [3,] 0.06646762
## [4,] 0.06221153
## [5,] 0.06259756
## [6,] 0.06442052
```

Evaluation

Uplifts

Bei den letzten Wählern im Datensatz sieht es anders aus. Diese hatten zuvor keine Botschaft erhalten und bekommen nun eine. Deswegen müssen wir hier mit -1 multiplizieren. Ergebnis:

```
uplift = matrix(d_hat - d_hat_reversed, 10000, 1)
#cause the second 5000 should now have a higher prob have to reverse signs
uplift[5001:10000] = -1* uplift[5001:10000]

head(uplift)

##            [,1]
## [1,] 0.05343989
## [2,] 0.05134148
## [3,] 0.06646762
## [4,] 0.06221153
## [5,] 0.06259756
## [6,] 0.06442052

tail(uplift)

##                [,1]
##  [9995,] 0.05097115
##  [9996,] 0.05402895
##  [9997,] 0.05538946
##  [9998,] 0.05282171
```

Outlook: Thoughts on Deployment

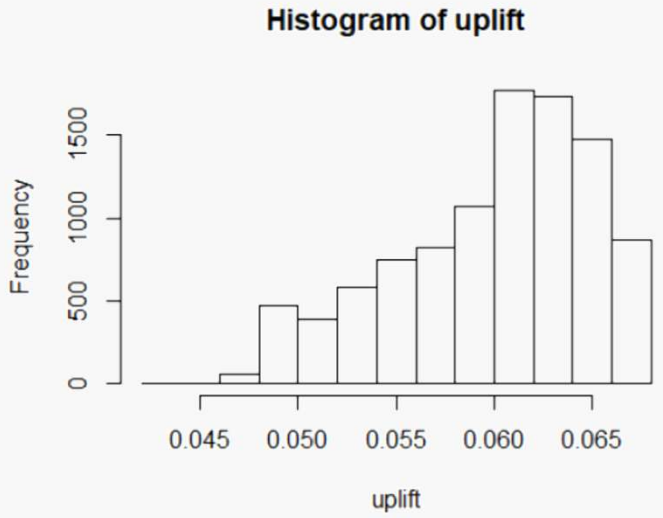
Auf welche Wähler würden Sie Ihre Botschaften konzentrieren?

Welche weiteren Schritte würden Sie im Modellbau vornehmen?

Evaluation

Uplifts

Man könnte also wie folgt vorgehen:




```
voterMatrix = matrix(cbind(uplift,dfrm_KUNDE$VOTER_ID),10000,2)

head(order(voterMatrix, decreasing = T))#should show most promising IDs first

## [1] 17846 17144 19982 13912 18209 13772
```

NOTE: Uplift Modeling refines/adds colour to the mailing case study of Larose! What would you change?