Verantwoord omgaan met onderzoekgegevens

Hoe je data- en syntaxbestanden transparant opslaat en je analyses repliceerbaar maakt

Damian Trilling

d.c.trilling@uva.nl
@damian0604

Afdeling Communicatiewetenschap
Universiteit van Amsterdam

Versie 0.1
Maart 2013

Waarom dit document?

Omdat. . .

- iedereen erbij gebaat is als onderzoek gerepliceerd kan worden.
- je bij oneenigheid ook later nog kan laten zien wat je gedaan hebt.
- het de reputatie van sociaalwetenschappelijk onderzoek in z'n algemeenheid ten goede komt als helder is hoe we tot onze resultaten komen.
- het je werk kan besparen.

(Dit document is specifiek bedoeld voor CW-afstudeerprojecten, maar kan uiteraard ook door anderen als algemene handleiding worden gebruikt.)

Uitgangspunten

Voordat je begint aan het lezen van dit document heb je het volgende al gedaan:

1. Hoe je aan je ruwe data bent gekomen is gedocumenteerd (en staat later in de methodesectie van je scriptie/paper).
2. Het meetinstrument (codeboeken, vragenlijsten, . . .) is opgeslagen.
3. Het onderzoeksmateriaal (artikelen, blogposts, tweets) eveneens.

Uitgangspunten

. . . en dit wordt hier uitgelegd:

1. hoe je het beste met verschilende versies van je dataset kunt omgaan
2. hercoderen en cleanen – maar wel verantwoord!
3. je analyses repliceerbaar maken – voor jezelf en voor vakgenoten

⇒ **Alles wordt gedocumenteerd, zo weinig mogelijk handmatige stappen!**

Datasets

Datasets

Wat opslaan?

1. de ruwe dataset (de data zoals ze binnenkomen)
2. een syntax die alle nodige hercoderingen uitvoert
3. een werkbestand om je analyses op los te laten

Op een centraal toegankelijke plek ($\Rightarrow$de dropbox-map), eventueel ook in een database of online beschikbaar stellen

Datasets

Waarom?

- Je kan fouten ongedaan maken (door de syntax aan te passen en opnieuw te runnen)
- Je hebt gedocumenteerd wat je precies hebt gedaan
- Je hebt geen tientallen bestanden (dataset-echtallerlaatsteversie-23.sav etc.) maar precies twee
- Anderen kunnen het je nadoen
- Je weet anders zelf over een tijdje ook niet meer hoe je tot je resultaten bent gekomen

1. De ruwe dataset

Wat?

- De data zoals ze binnen zijn gekomen (van Qualtrics, na het invoeren van inhoudsanalysedata, . . .)
- Dit bestand wordt NEVER NOOIT aangepast of bewerkt.

In dit afstudeerproject: één bestand per groepje

2. De hercoderingssyntax

Wat?

- De syntax opent het originele bestand. . .
- . . . voert vervolgens alle nodige aanpassingen door (hercoderingen, het aanmaken van schalen, invoerfouten corrigeren) . . .
- . . . en slaat het resultaat onder een andere naam op.

Tip: Maak gebruik van comments om de syntax leesbaar te houden!

In dit afstudeerproject: Of één per groepje, of één per persoon. Ik raad het eerste aan.

3. Het werkbestand

Wat?

- Dit is het resultaat van je hercoderingssyntax.
- Hierop draai je al je analyses.

Hercoderen en cleanen – maar wel verantwoord

Een voorbeeld

```
546  /* EXTERNAL EFFICACY*/.
547  RECODE V118_6 (7=1) (6=2) (5=3) (4=4) (3=5) (2=6) (1=7) INTO V118_6r.
548  RECODE V118_7 (7=1) (6=2) (5=3) (4=4) (3=5) (2=6) (1=7) INTO V118_7r.
549  RECODE V1181_1 (7=1) (6=2) (5=3) (4=4) (3=5) (2=6) (1=7) INTO V1181_1r.
550  COMPUTE exteff = (V118_6r + V118_7r + V1181_1r)/3 .
551  EXECUTE .
552  VARIABLE LABELS exteff 'External political efficacy'.
553
554  /* POLITICAL CYNICISM */.
555  RECODE V1181_2 (7=1) (6=2) (5=3) (4=4) (3=5) (2=6) (1=7) INTO V1181_2r.
556  RECODE V1181_3 (7=1) (6=2) (5=3) (4=4) (3=5) (2=6) (1=7) INTO V1181_3r.
557  COMPUTE pcyn = (V1181_2r + V1181_3r)/2 .
558  EXECUTE .
559  VARIABLE LABELS pcyn 'Political cynicism'.
```

De bestanden origineel.sav, recode.sps en werkbestand.sav vind je in de dropbox. Het gaat om een grootschalig surveyonderzoek waaraan drie onderzoeks hebben meegewerkt.

Hoe doe je het?

De belangrikste commando's

- RECODE
- COMPUTE
- IF ((id=34) OR (id=22)) AND (gender=1) V1=23.

Behalve het laatste (IF) kan je dit ook allemaal via de menus doen (en dan op "Paste" ipv "OK" klikken).

Cases verwijderen

- Soms is het noodzakelijk cases niet mee te nemen in je analyse. Maar let op: *Je moet goet kunnen verantwoorden waarom je cases verwijdert!* Dit is een *slippery slope* naar *sloppy science*! Je moet je keuze in je scriptie héél goed uitleggen!
  (Voorbeeld: Je hebt gemeten hoe lang mensen in een experiment het stimulusmateriaal hebben gelezen en het blijkt dat sommigen meteen hebben doorgeklikt)
- **Als je cases verwijdert, dan gebeurt dat nooit handmatig, maar altijd in de syntax, zodat het ongedaan kan worden gemaakt en duidelijk is WELKE cases er precies zijn verwijderd en op basis van welke criteria**
- SELECT IF (leesduur>5) /* korter dan 5 sec gelezen weg*/.
- SELECT IF NOT (id=125) /* case nummer 125 weg */.

Repliceerbare analyses

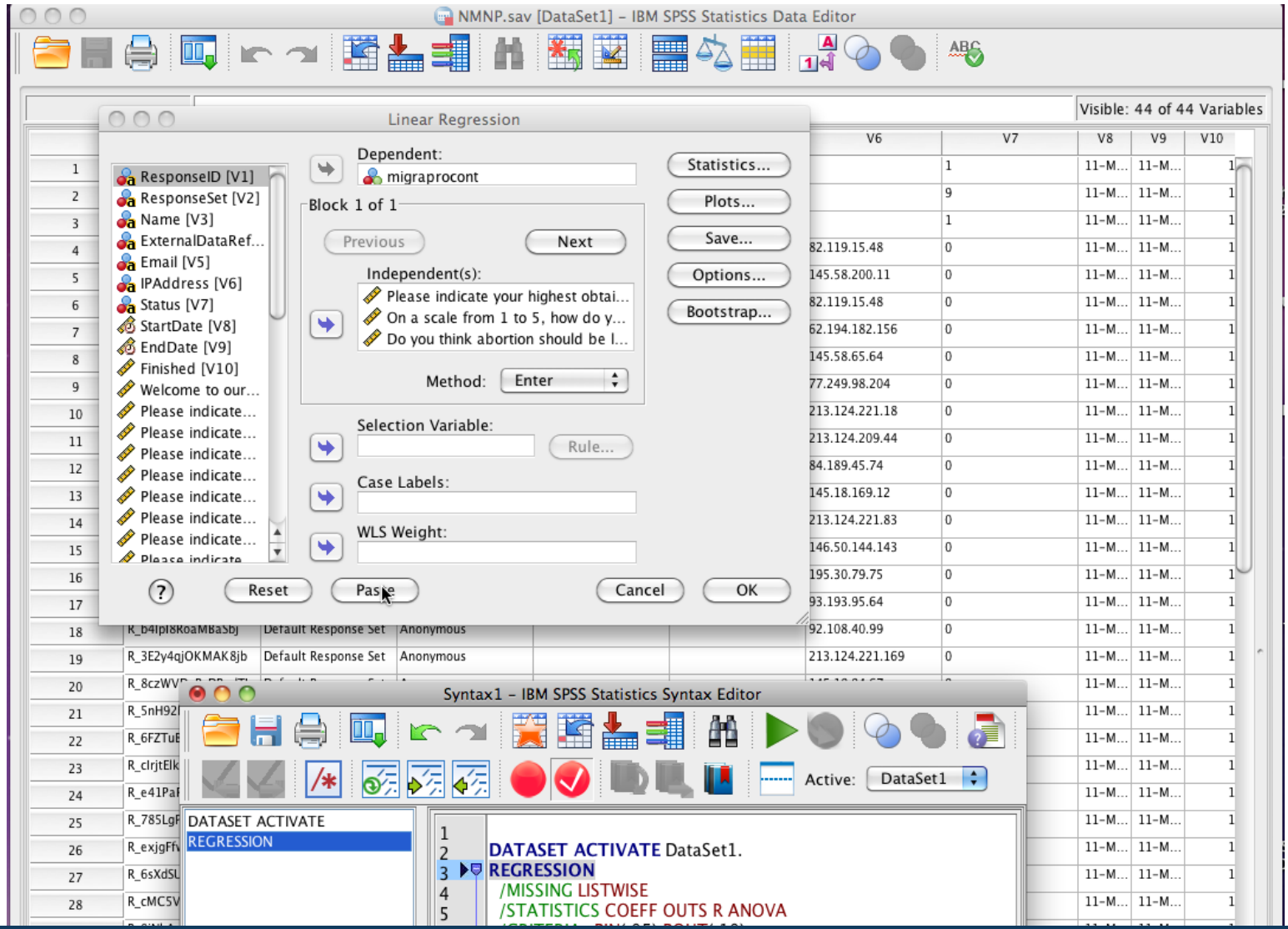
Repliceerbare analyses

Syntax, syntax, syntax!

- Je zou op een blad papier kunnen opschrijven welke analyses je precies hebt gedraaid
- Maar makkelijker is het om dat met een syntax te doen
- Voordeel: Als je later iets wilt aanpassen (een andere variabele meenemen, . . .) dan is dit met een muisklik gedaan!
- En: Je kan dezelfde analyses nog een keer op een andere dataset loslaten.

In dit afstudeerproject: één syntax per student, waarin alle in je scriptie gemaakte analyses staan, plus eventueel aanvullende analyses.

Syntax invoegen



Tot slot

Kort samengevat:

Je doet niks wat niet gedocumenteerd is. Je gaat niet handmatig in datasets data aanpassen. Je hebt een ruwe databestand, een hercoderingssyntax, een werkbestand en een aantal analysesyntax-betanden. Je stelt deze bestanden beschikbaar in de dropbox. Mocht er iets onduidelijk zijn, vraag je het aan mij.

Checklist afstudeerproject #dbdb

De volgende dingen staan in de dropbox:

1. een beschrijving van de dataverzameling (in je scriptie)
2. het codeboek
3. een logboek dat tijdens het coderen is bijgehouden
4. het onderzoeksmateriaal (artikelen, blogposts, tweets)
5. datasets en syntaxbestanden van pretest(en) en intercodeurbetrouwbaarheidstest(en)
6. de ruwe data
7. de hercoderingssyntax (of, in het geval dat je bijvoorbeeld een python-script hebt gebruikt, het script plus een uitleg)
8. het werkbestand
9. de analysesyntax

Vragen of opmerkingen?

Damian Trilling

d.c.trilling@uva.nl
@damian0604