Intro to SAS

Hadley Wickham

Outline

- Transformation and subsetting
- Linear models
  - Revision
  - PROC GLM

Procedures

- Procedures in SAS are the equivalent of functions in R (except they generally do a lot more)
- We'll learn our first procedure: PROC PRINT
- PROC PRINT; run;
- PROC PRINT data=XXX; run;

Your turn

- Load the server survey data into SAS
- Use proc print to print it out
- Read the help for proc print (SAS Products | Base SAS | SAS Procedures | Procedures | The PRINT procedure) and only print out one or two variables
- Experiment with the other options

Selecting variables

- Variables in dataset: c b a x1 x2 x3 age_old age_young
  - x1-x3 = x1 x2 x3
  - age: = age_old age_young
  - c--a = c b a
  - c--age_young = all variables

Data in SAS

- R processes data by columns, SAS processes by rows
- In the data step we can do all sorts of complicated things, but we're just going to stick with simple variable transformations

```
data ss;
set ss;
age = 2007 - birth_yr;
lifeworked = yrs_experience / age;

group = "unknown";
if (age ^= . & age <= 20) then group = "teenage";
if (age > 20 & age <= 30) then group = "20s";
if (age > 30 & age <= 50) then group = "middle-aged";
if (age > 50) then group="elderly";

run;
```

Your turn

- Load the server survey data into SAS
- Clean up the sex variable like you did for the project
- Clean up percent tip
- Practice creating any other transformations that you used

Subsetting

- Similar to R
- Three ways:
  - Create new data set
  - Use subset in procedure x2

```
data teenagers;
set ss;
where group = "teenage";
run;

proc print data=teenagers;
run;
```

```
proc print data=ss;
where group eg "teenage";
run;

proc print data=ss(where=(group = "teenage"));
run;
```

Differences from R

- ^= instead of !=
- x in (1 4 5) instead of x %in% c(1,4,5)

Your turn

- Practice your subsetting!

Linear models

- How much do you remember?
- Explain/summarise the pattern in a single "y" by any number of "x" variables
- $Y \sim \text{Normal}(X\beta, \sigma I)$

Special cases

- Used for both continuous and categorical variables

- t-test
- ANOVA
- ANCOVA

Assumptions

- Errors normally distributed
- Errors independent
- Errors have constant variance
- Model is correct (linear)

PROC GLM

- Help: SAS Products | SAS/STAT | SAS/STAT users guide | The GLM Procedure
- PROC GLM data= ;
- MODEL y = a b c;
- RUN;

Your turn

- Predict percent tip from server demographics
- Predict percent tip from server behaviours
- What output do you get?
- Which is more useful?
- Which variables are useful at predicting tip?

Continuous vs categorical

- Should you fit those explanatory variables as continuous or categorical?
- Why does it matter? What's the difference?
- SAS doesn't remember the type of variables like R does, so you have to specify categorical variables every time using the CLASS statement

Your turn

- Refit the model using categorical variables
- What difference does it make?
- Does the model fit better?

More useful output

- LSMEANS varname / pdiff;
- LSMEANS varname / pdiff adjust=Tukey;

Adding interactions

- y = a b a*b
- y = a|b
- y = a b c a*b b*c a*c a*b*c (y = a|b|c)
- y = a b c a*b b*c a*c (y = a|b|c@2)

Extending the model

- What if we want to examine these effects separately for men and women - what could we do?
-

Extending the model

- two models, use subsetting
- use the by statement (+ proc sort or / unordered)
- use interactions

Your turn

- Use one of your models to compare if there are different male and female effects