Introduction to $\mathcal{R}$

Session 6: GLMs

Dag Tanneberg[1]

Potsdam Center for Quantitative Research
University of Potsdam, Germany
October 11/12, 2018

[1]Chair of Comparative Politics, UP, dag.tanneberg@uni-potsdam.de

Introduction

Before we start...

- Quit & reopen RStudio.
- Load "./06/dta/asoiaf.csv"" from the course material.
  - **Remember**: Uncheck the option "Strings as factors"
- Open a new script file.
- Execute the following code:

```
asoiaf[, "died"] <- !is.na(asoiaf[, "book_of_death"])
```

- Install the "car" package.

What do we intent to do?

- **Question**: What's the chance that Jon Snow is going to die?
- **Means**: Regression on a linear combination of predictors

$$p(Death = 1|\mathbf{X}, \beta) = \beta_{\mathbf{0}} + \sum_{\mathbf{K}} \beta_{\mathbf{k}} \mathbf{x_k}$$

- **Problem**: Chance of death is not a well-behaved response.
  a. We don't obseverve probabilities but discrete events.
  b. Probabilities are restricted to $[0, 1]$, but $\mathbf{X}\beta$ can take any value.
- **Challenge**: Map the linear combination $\mathbf{X}\beta$ into a domain which fits our response.

Some Intuition on GLMs

- Applies to many quantities of interest, e.g.,
  - Household income
  - Satisfaction with democracy
  - Number of bills per session of parliament
  - ...
-

Outline

The Basics of Running GLMs in $\mathcal{R}$

Generic Format of Fitting GLMs

```
fit <- glm(
  formula = <formula>,
  family = <family>(link = "<link>"),
  # Defaults to gaussian(link = "identity"). Therefore
  # we skip the lm() function and OLS.
  data = <data>,
  weights = <weights>, # Be careful! Meaning changes
                       # depending on <family>.
  subset = <subset>,
  na.action = na.omit, # Retains only complete cases.
  <...> # Options to tweak the optimizer.
)
```

$\mathcal{R}$'s Formula Interface[2]

Generic Example

$$y \sim x_1 + x_2 + \cdots + x_k$$

Formula Creation

Symbol	Meaning	Example
:	Specify an interaction	$y \sim x : z \Rightarrow y = xz$
$*$	Specify all possible interactions	$y \sim x * z \Rightarrow y = x + z + xz$
$\widehat{}$	Specify interactions up to some degree	$y \sim (x + z)^2 \Rightarrow y = x + z + xz$
.	Wildcard for all other variables	$y \sim . \Rightarrow y = x + z + w + \ldots$
-	Remove variable(s)	$y \sim (x + z)^2 - x : z \Rightarrow y = x + z$
-1 OR $0+$	Remove the intercept	$y \sim x - 1$ OR $y \sim 0 + x$
$I()$	Arithmetical transformation	$y \sim I(x^2) \Rightarrow y = x^2$
function	Other mathematical transformations	log10$(y) \sim x \Rightarrow log_{10}(y) = x$

[2]Adapted from Kabacoff, R. 2011. R in Action. Shelter Island: Manning Publications, p. 178.

$\mathcal{R}$'s Formula Interface, contd.

Exercise How would you write the following formulas?

1. $y = x + z + xz$
2. $y = x + x^2 + x^3$
3. $log_e(y) = x + z + w + xz + xw + wz$
4. y as a function of variables in the data but k

Family Generators and Link Functions in glm()[3]

A Practical Example

```
glm(<...>, family = binomial(link = "logit"), <...>)
```

	link = "<arg>"							
	μ	μ^{-1}	$ln(\mu)$	$ln(\frac{\mu}{1-\mu})$	$\Phi(\mu)$	$ln[-ln(1-\mu)]$	$\sqrt{\mu}$	$\frac{1}{\mu^2}$
family	identity	inverse	log	logit	probit	cloglog	sqrt	1/mu^2
gaussian()	●	○	○					
binomial()			○	●	○	○		
poisson()	○		●				○	
Gamma()	○	●	○					
inverse.gaussian()	○	○	○					●
quasi()	●	○	○	○	○	○	○	○
quasibinomial()				●	○	○		
quasi()	○		●				○	

Legend: ● default, ○ possible

[3] Adapted from Fox, J. and S. Weisberg. 2011. An R Companion to Applied Regression. 2nd ed. London: SAGE, pp. 231, 233.

Get Your Hands Dirty

Now it's your turn. Fit a

- **logistic** regression model which
- predicts **died**
- from **allegiances**,
- the full interaction of **gender** and **nobility**,
- a cubic polynomial on **age_in_chapters**,
- and save it to an object called **myfit**.

Working With Regression Results

Testing Assumptions