Generalized linear regression

Outline

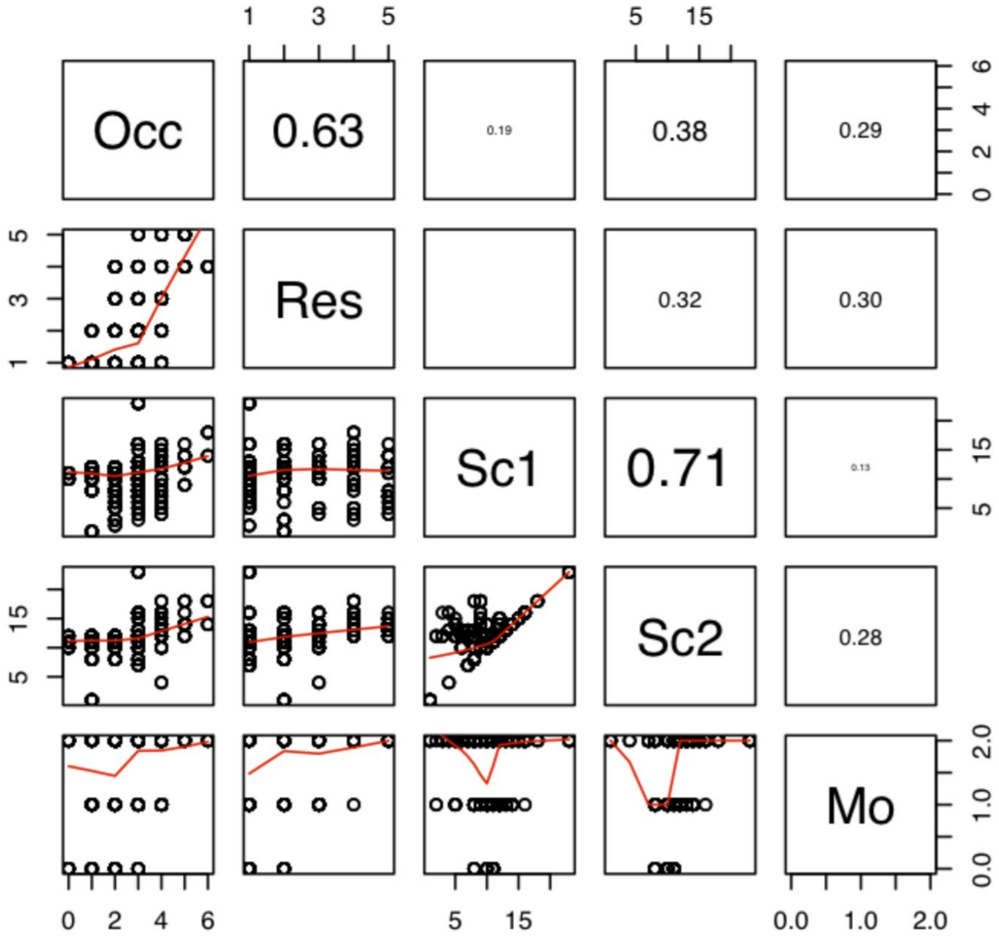
- Homework tips
- Interactions between independent variables
- *Generalized linear regression*
  - In particular, *logistic regression*, used for binomial dependent variables
- For home consumption: isotonic regression

Homework 05 tips (1/)

- If you're using `stopifnot` to verify properties, and the condition fails, try `print`ing out the relevant values just before you run `stopifnot`.
    - Unfortunately R doesn't let you specify a logging statement with `stopifnot`, though Python does with `assert`.
- The command `all.equal` can be used to check if two values are very close, and plays nicely with `stopifnot`.

Homework 05 tips (2/)

- A few people were confused by the notation `Y ~ 1`:
  - If you have `Y ~ X` and you want to "drop out" *X*, R doesn't let you write `Y ~`.
  - The `1` here symbolizes the intercept.
  - If for some reason you want a model *without* an intercept you can write `Y ~ -1`.
- Generalizing a bit, if you have `Y ~ X1 + X2`:
  - If you want to "drop out" X_1, you write `Y ~ X2`.
  - If you want to "drop out" X_2, you write `Y ~ X1`.
- The function `lrtest` from the `lmtest` package can compute the log-likelihood ratio test (both the test statistic and the *p*-value) for two models so long as one nests the other; you still have to fit the "dropped out" models though.



Occ
Res
Sc1
Sc2
Mo
0.63
0.19
0.38
0.29
0.32
0.30
0.71
0.13
0.28

Questions?

Interactions

Interaction terms

In some cases we are interested not just in the effect of a given independent variable (IV) but its *interaction* with another IV.

The nature and interpretation of the interaction depends on the the types of IVs involved.

Types of interaction

- Interaction of two binomial IVs: is there an change in *Y* when both X_1 *and* X_2 are active (i.e., true) beyond that associated with X_1 and X_2?
  - Are the effects of X_1 and X_2 on *Y* independent?
- Interaction of a binomial IV X_b and a continuous IV X_c: is the change in *Y* associated with X_c different when $X_{c'}$, is active?
  - Is the slope of $X_{c,}$ with respect to *Y* different when X_1 is active?
- Interaction of two continuous IVs: is *Y* also sensitive to the product of X_1 and X_2?

The interaction of two multinomial IVs is best understood by decomposing them into binomial IVs.

Specifying interactions in R formulae

Long form:

```
Y ~ X1 + X2 + X1:X2
```

Short form:

```
Y ~ X1 * X2
```

You *can* specify an interaction without a main term (e.g., `Y ~ X1:X2`) but it is rarely needed.

Example (1/)

Is petal length influenced by the *interaction* between sepal length and sepal width?

Q: What kind of interaction is this?

A: An interaction between continuous IVs.

```
> r.simple <- lm(Petal.Length ~ Petal.Width +
+                Sepal.Length + Sepal.Width,
+                data = iris)
> summary(r.simple)
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.26271    0.29741  -0.883    0.379
Petal.Width   1.44679    0.06761  21.399   <2e-16 ***
Sepal.Length  0.72914    0.05832  12.502   <2e-16 ***
Sepal.Width  -0.64601    0.06850  -9.431   <2e-16 ***
...
```

```
> r.intrct <- lm(Petal.Length ~ Petal.Width +
+                Sepal.Length * Sepal.Width,
+                data = iris)
> summary(r.intrct)
...
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.71482    1.56623   0.456   0.6488
Petal.Width            1.43584    0.06991  20.539   <2e-16 ***
Sepal.Length           0.56175    0.26970   2.083   0.0390 *
Sepal.Width           -0.97041    0.51486  -1.885   0.0615 .
Sepal.Length:Sepa..    0.05642    0.08874   0.636   0.5259
...
```

Example (2/)

The three-way interaction model

```
Petal.Length ~ Petal.Width + Sepal.Length + Sepal.Width +
               Petal.Width:Sepal.Length +
               Petal.Width:Sepal.Width +
               Sepal.Length:Sepal.Width +
               Petal.Width:Sepal.Length:Sepal.Width
```

can also be fit (if you have enough data), but is basically uninterpretable.

Example (3/)

Is there an interaction between age and gender in the ANAE low back merger data (from homework 05)?

Q: What kind of interaction is this?

A: An interaction between a continuous IV (age) and a binomial IV (gender...uh... at least as it is coded in that data).

```
> r.simple <- lm(distance ~ age + gender + dialect,
+               data = anae)
> summary(r.simple)
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  176.8756     7.8732  22.465  < 2e-16 ***
age            0.7735     0.2954   2.618 0.009177 **
gender1        4.4896     4.4425   1.011 0.312821
dialect1    -103.6565    30.3734  -3.413 0.000708 ***
dialect2    -138.7272    34.7547  -3.992 7.80e-05 ***
dialect3    -132.0029    18.7318  -7.047 8.00e-12 ***
dialect5       1.5156    48.6868   0.031 0.975181
...
```

```
> r.intrct <- lm(distance ~ age + gender + age:gender +
+                dialect, data = anae)
> summary(r.intrct)
...
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    177.0400     7.8752  22.481  < 2e-16 ***
age              0.6877     0.3080   2.233 0.026093 *
gender1          4.2793     4.4478   0.962 0.336575
age:gender1      0.2984     0.3024   0.987 0.324383
dialect1      -102.7246    30.3890  -3.380 0.000795 ***
dialect2      -141.1245    34.8407  -4.051 6.14e-05 ***
dialect3      -132.4366    18.7376  -7.068 7.02e-12 ***
...
```

Example (4/)

There is no consensus whether it is sensible to test for interactions (e.g., of age and gender) when one or the other non-interaction term is non-significant (e.g., above, where the standard error for age was larger than the coefficient).

Nota bene

- Caution is necessary when "dropping" (i.e., doing likelihood ratio tests on) models with interaction terms:
  - If the model is `Y ~ X1 * X2`, to drop the interaction you write `Y ~ X1 + X2`.
  - To make this even clearer, you can specify the full model as `Y ~ X1 + X2 + X1:X2`.
  - If you drop a (non-interaction) independent variable, you should also remove its interactions.
- `drop1` does not understand interactions and gives nonsensical results if they are present; stepwise fitting functions may similarly be confused.
- Three- and four-way interactions are hard to interpret, burn up degrees of freedom quickly, and tend to give short shrift to main effects:

  "Analysts usually steer clear of higher-order interactions".

Generalized linear regression

Generalizing linear regression

Two key assumptions of linear regression (as well as ANOVA) is that the dependent variable (DV) is

- normally distributed (or the central limit theorem applies), and
- a linear sum of the coefficients and their IVs.

This assumption is *flagrantly* violated when the DV is a proportion or probability, as

- probabilities violate the assumption of homogeneity of variance, and
- for probability DVs a linear model can predict $p < 0$ or $p > 1$.

Problems with percentages

A change of a percentage $\hat{p} = .5$ is "less" (according to the binomial distribution) of a change than a change for probability close to 0 or 1, so

- effects close to 0 or 1 are underestimated and
- effects close to .5 are overestimated.

"In what space can we capture these intuitions?"

Desiderata:

- smooth, continuous, differentiable transformation function
- domain $[0, 1]$, range $(-\infty, +\infty)$

The arcsine transformation

One traditional answer to these questions is the *arcsine* (or *angular*) transformation, defined by the inverse sine of the square root of the proportion, or

$$\arcsin \sqrt{[p]}$$

which has a range of [0, 2π]. Or in R:

```
> asin(sqrt(p))
[1] 0.4636476
```

From probabilities to odds

The odds of a probability p is simply:

$$O = p \,/\, (1 - p)$$

This has the range [0, +∞].

For instance, for $p = .9$, $O = 9$, and for $p = .1$, $O = 0.1111$.

From odds to log-odds

Because of the strange range…

e.g., $p < .5$ implies $0 < O < 1$, whereas p > .5 implies $1 < O < \infty$,

it is often preferable to work in log-space, where the range is $[-\infty, +\infty]$.

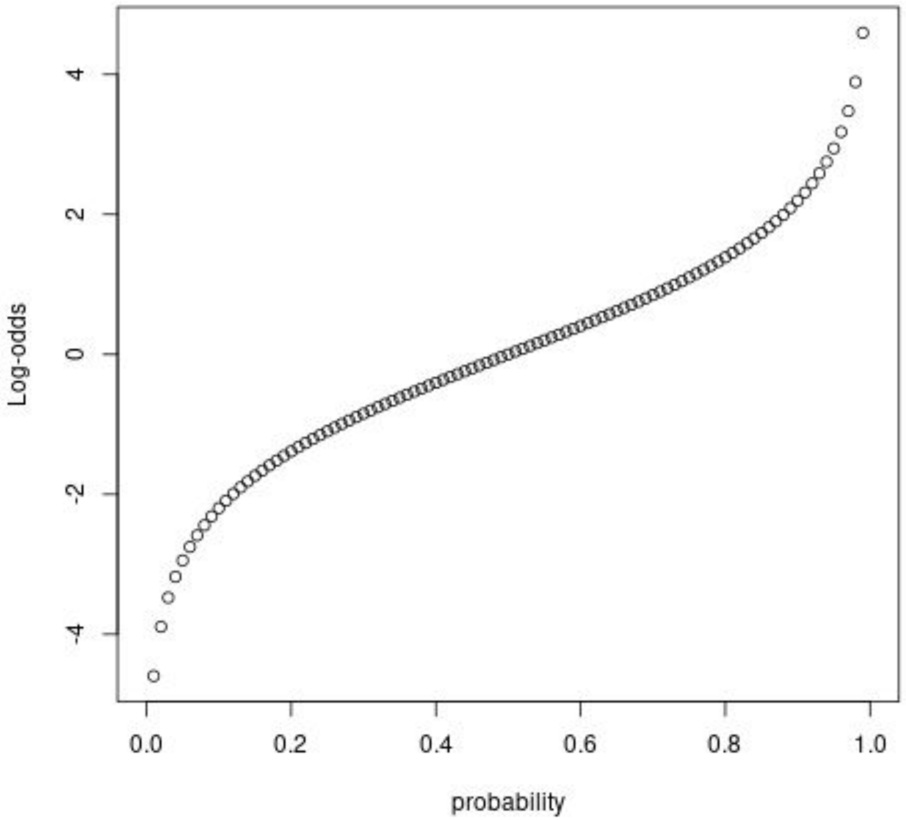
$$\begin{aligned} \log O &= \log p - \log(1 - p) \\ &= \log c - \log(N - c) \end{aligned}$$

For instance, for $p = .9$, $\log(O) = 2.197$, and for $p = .1$, $\log(O) = -2.197$.

Introducing `qlogis`

In R, the transformation from probabilities to log-odds, the *logit,* is performed by `qlogis`.

```
> qlogis(seq(0, 1, .1))
 [1]       -Inf -2.1972246 -1.3862944 -0.8472979 -0.4054651
 [6]  0.0000000  0.4054651  0.8472979  1.3862944  2.1972246
 [7]  Inf
```



Log-odds
4
2
0
-2
-4
0.0
0.2
0.4
0.6
0.8
1.0
probability

Introducing logistic regression

The framework of *generalized linear models* are linear models augmented with *link functions* (such as the logit) which map arbitrary types of DVs onto a linear function.

With some magic (not covered in this class), we can then estimate the parameters of such models.

A generalized linear models with logit link functions are known as *logistic regression* models.

FYI: the *logistic* is the inverse of the logit function, mapping from log-odds to probabilities.

Logistic regression in R

To specify a logistic regression, we use the function `glm` (which fits generalized linear models) and specify `family = binomial` (which enables a logit link function, giving us *logistic* regression) in particular.

Nearly all other linear model functions work the same; we can call `summary`, compute `residuals`, perform the likelihood ratio test with `drop1`, etc.

An example from the [ʃ]treets of Columbus (1/)

The envelope of variation is pronunciation of word-initial *str*- as [str] vs. [ʃtr].

Data collected using a *rapid anonymous* design: ask for directions to a nearby bank so as to elicit tokens of *street* (cf. Labov 1966 on post-vocalic *r* in New York via *fourth floor*, Prichard 2010 on /ay/-monophthongization in Atlanta via *five o'five*).

An example from the [ʃ]treets of Columbus (2/)

Predictors include:

- Gender
- Emphasis (normal vs. a second rendition after "what did you say?")
- Age (coded as "young", "middle", or "old")
- Social class (coded as "working class", "lower middle class", and "upper middle class")

```
> xtabs(~ str + emphatic, data = cbus)
     emphatic
str    Less More
  shtr   43   14
  str    77  106
```

```
> r <- glm(str ~ emphatic, data = cbus, family = binomial)
> summary(r)

Call:
glm(formula = str ~ emphatic, family = binomial, data = cbus)
...
Coefficients:

               Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.5826     0.1904   3.060  0.00221 **
emphaticMore     1.4418     0.3422   4.213 2.52e-05 ***
...
```

Interpretation notes

Using the standard procedure of computing estimates for *Y* by adding the intercept and the product of coefficients and IVs, we obtain a number in log-odds space. We can convert back to an estimated probability using the R function `plogis`, the inverse of `qlogis`.

E.g., to estimate *P*(str | "more emphatic"), we have:

```
> intercept <- 0.5826
> moreEmphatic <- 1.4418
> plogis(intercept + moreEmphatic)
[1] 0.8833352
```

Nota bene

There is a strong connection between (binomial) logistic regression for statistical inference and so-called *multinomial logistic regression* (or *maxent models*) used for classification in speech and language processing, though

- they often use different representations of the IVs (e.g., dense continuously-valued vs. sparse booleans), and
- they often use different learning algorithms (e.g., iteratively-reweighted least squares vs. stochastic gradient descent).

Questions? Please take them to email, or Slack.