ADVANCED BAYESIAN MODELING

Shakespeare Plays: Data and Model

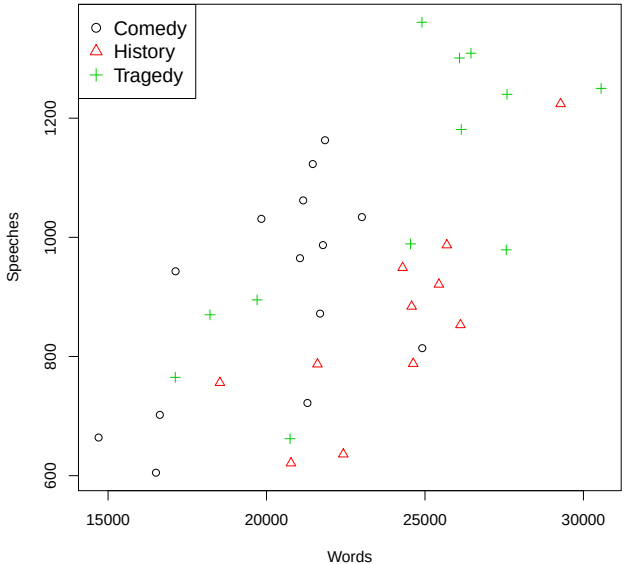
For William Shakespeare's 37 undisputed plays, let

$$y_i \quad = \quad \text{number of speeches in play } i$$

where a "speech" is either a unit of dialogue spoken by a character, or a stage direction.

We will model number of speeches conditionally using

- Traditional genre classification (comedy, history, tragedy)
- Total number of words


Comedy
History
Tragedy
Speeches
600
800
1000
1200
15000
20000
25000
30000
Words

```
> ss <- read.csv("shakespeare.csv", comment.char="#")

> head(ss)
                          Composed   Genre Words Speeches Scenes Characters
All's Well That Ends Well     1602  Comedy 23009     1034     23         26
Antony and Cleopatra          1606 Tragedy 24905     1361     42         57
As You Like It                1599  Comedy 21690      872     22         27
Comedy of Errors              1589  Comedy 14701      664     11         19
Coriolanus                    1607 Tragedy 27589     1240     28         60
Cymbeline                     1609 Tragedy 27565      979     27         40
```

Data compiled from Open Source Shakespeare (`http://www.opensourceshakespeare.org`).

Tragedies have the greatest mean number of speeches:

```
> with(ss, by(Speeches, Genre, mean))
Genre: Comedy
[1] 906.2143
------------------------------------------------------------
Genre: History
[1] 855.0909
------------------------------------------------------------
Genre: Tragedy
[1] 1066.833
```

But this does not account for overall length of the plays (e.g., tragedies tend to be longer than comedies).

Linear Regression

$$y_i \mid \theta, X \sim \text{indep. } \mathrm{N}(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}, \sigma^2) \qquad i = 1, \ldots, 37$$

i.e.,

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

where

x_{i1} = indicator that play i is a comedy

x_{i2} = indicator that play i is a history

x_{i3} = indicator that play i is a tragedy

x_{i4} = *standardized* total number of words in play i

Notice: No explicit intercept variable, but intercept is present implicitly –

$$x_{i1} + x_{i2} + x_{i3} = 1 \qquad \text{for all } i$$

Indicator variables that implicitly define the intercept should not be standardized.

The total words variable x_{i4} is standardized – reasonable because it does not involve the intercept, and its values are large in magnitude.

β_1, β_2, and β_3 represent expected number of speeches in each genre, after adjustment for total number of words. For example,

$$\beta_1 \quad = \quad \text{expected \# speeches for a comedy having overall average words}$$

(average number of words for all plays, because x_{i4} is standardized)

To compare, e.g., comedies to tragedies, after adjusting for total number of words, use $\beta_1 - \beta_3$.

β_4 represents expected increase in speeches (irrespective of genre) when total words increases by one standard deviation (over all plays).

Prior

We choose semi-conjugate priors

$$\begin{aligned} \beta \mid X &\sim \mathrm{N}(0,\, \sigma_\beta^2 I) \\ \sigma^2 \mid X &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

Choice of $\sigma_\beta^2 I$ is arbitrary, but should not matter if σ_β^2 is large enough.

Follows that

$$\beta_1, \beta_2, \beta_3, \beta_4 \mid X \sim \text{iid } \mathrm{N}(0,\, \sigma_\beta^2)$$

Let σ_β^2 be large and ν_0 small to make prior nearly noninformative (no prior info).