Statistical Methods for Insurance: Generalised Linear Models

Di Cook & Souhaib Ben Taieb, Econometrics and Business Statistics, Monash University

W7.C2

Generalised linear models

- Overview
- Types
- Assumptions
- Fitting
- Examples

Overview

- GLMs are a broad class of models for fitting different types of response variables distributions.
- The multiple linear regression model is a special case.

Three components

- Random Component: probability distributin of the response variable
- Systematic Component: explanatory variables
- Link function: describes the relaionship between the random and systematic components

Multiple linear regression

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad or \quad E(Y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Random component: y_i has a normal distribution, and so $e_i \sim N(0, \sigma^2)$
- Systematic component: $\beta_0 + \beta_1 x_1 + \beta_2 x_2$
- Link function: identity, just the systematic component

Poisson regression

$$y_i = exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2) + \varepsilon$$

- y_i takes integer values, 0, 1, 2, ...
- Link function: $ln(\mu)$, name=`log`. (Think of μ as $\hat{y}$.)

Bernouilli, binomial regression

$$y_i = \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)} + \varepsilon$$

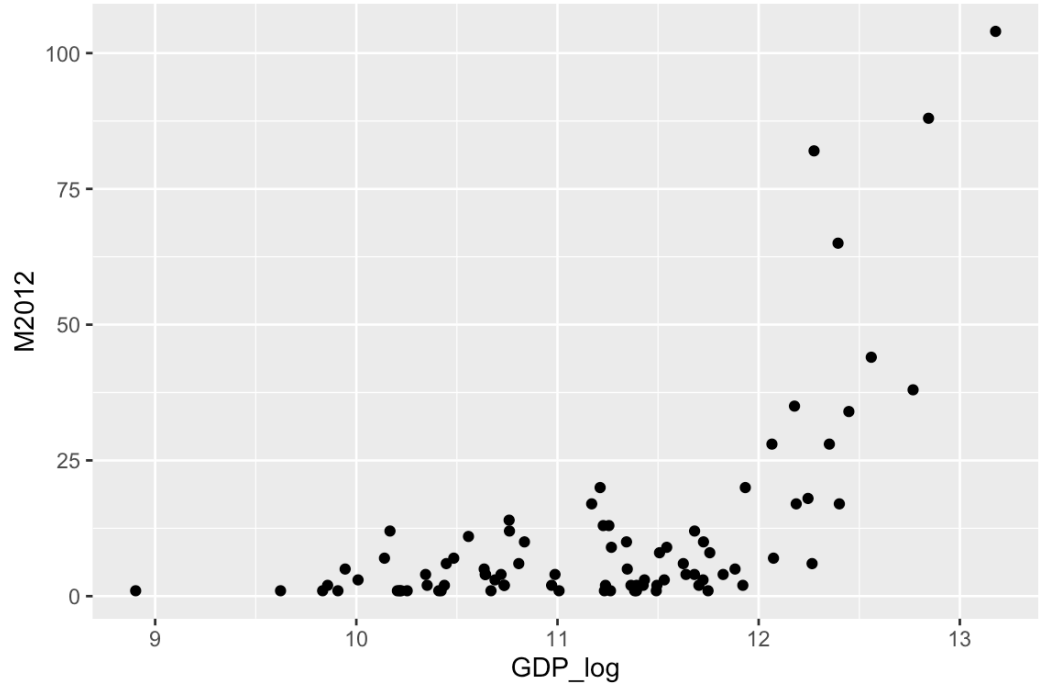
- y_i takes integer values, $\{0, 1\}$ (bernouilli), $\{0, 1, \ldots, n\}$ (binomial)
- Let $\mu = \frac{exp(\beta_0+\beta_1 x_1+\beta_2 x_2)}{1+exp(\beta_0+\beta_1 x_1+\beta_2 x_2)}$, link function is $ln\frac{\mu}{1-\mu}$, name=`logit`

Assumptions

- The data $y_1, y_2, \ldots, y_n$ are independently distributed, i.e., cases are independent.
- The dependent variable y_i does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal,…)
- Linear relationship between the transformed response (see examples below)
- Explanatory variables can be transformations of original variables
- Homogeneity of variance does NOT need to be satisfied
- Uses maximum likelihood estimation (MLE) to estimate the parameters
- Goodness-of-fit measures rely on sufficiently large samples

Example: Olympics medal tally

- Model medal counts on log_GDP
- Medal counts = integer, suggests use a Poisson model.



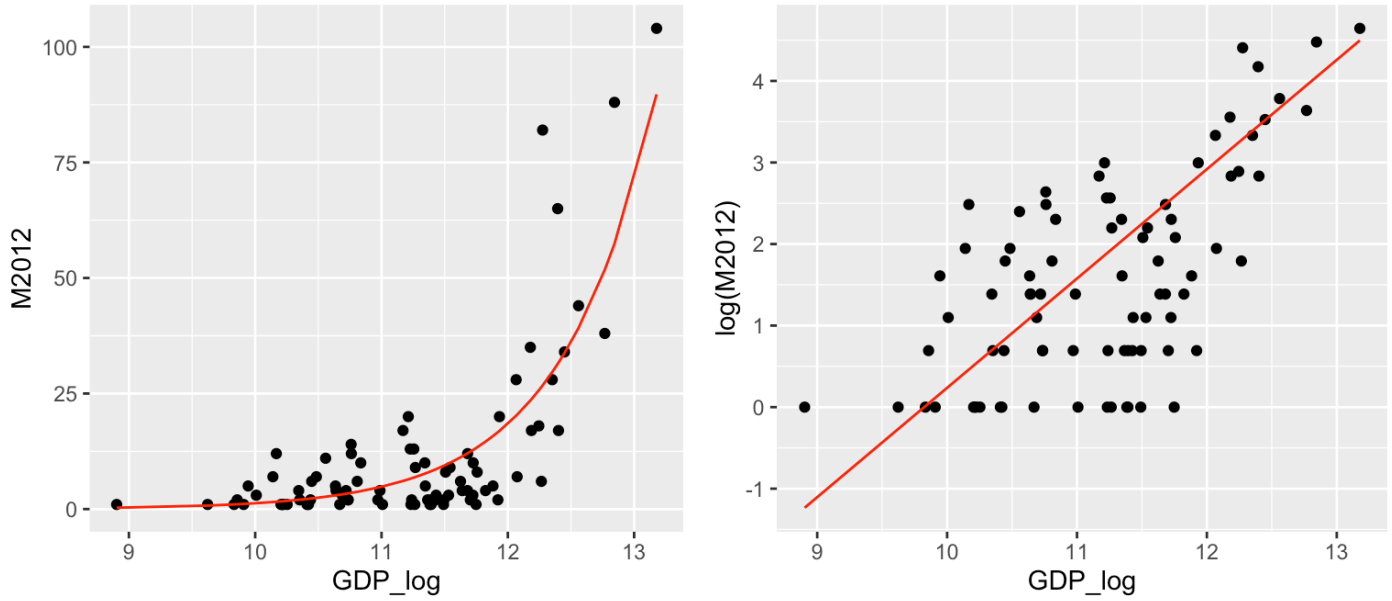
```
oly_glm <- glm(M2012~GDP_log, data=oly_gdp2012,
               family=poisson(link=log))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-13.2	0.54	-24	0
GDP_log	1.3	0.04	30	0

Your turn

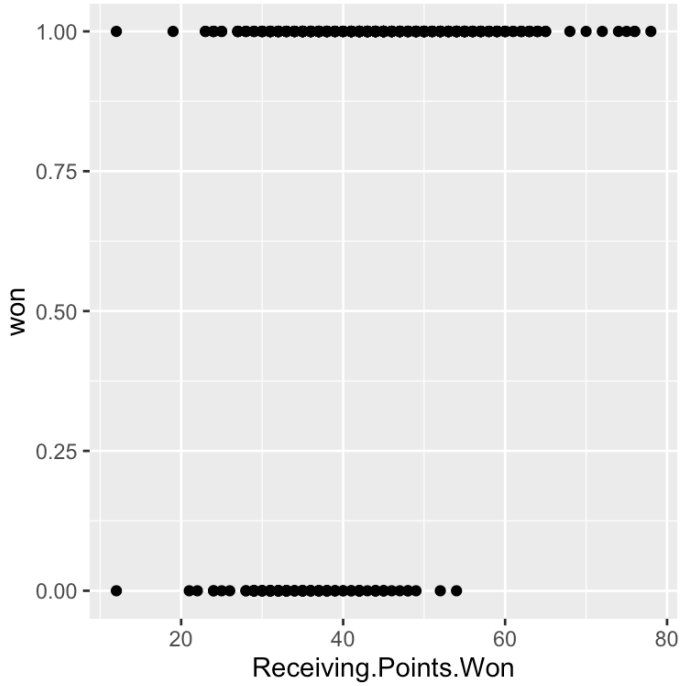
Write down the formula of the fitted model.

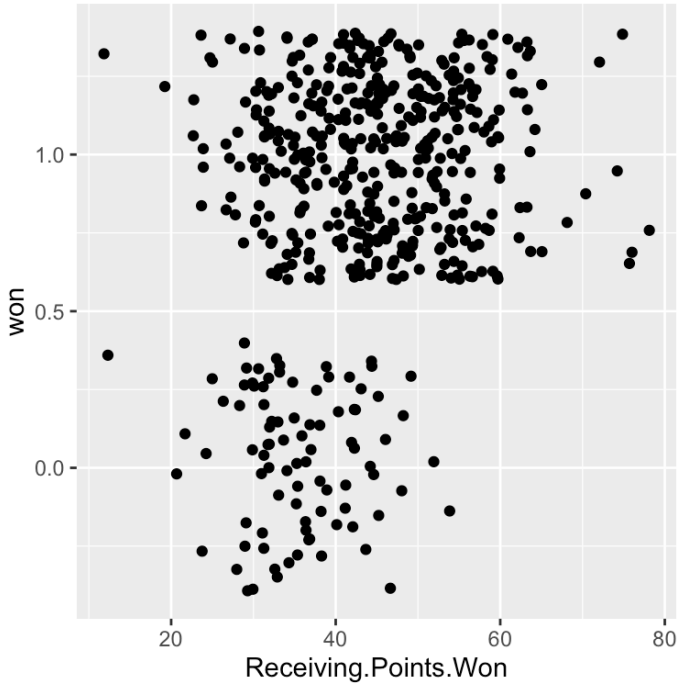
What does this model look like?



Example: winning tennis matches

We have data scraped from the web sites of the 2012 Grand Slam tennis tournaments. There are a lot of statistics on matches. Below we have the number of receiving points won, and whether the match was won or not.





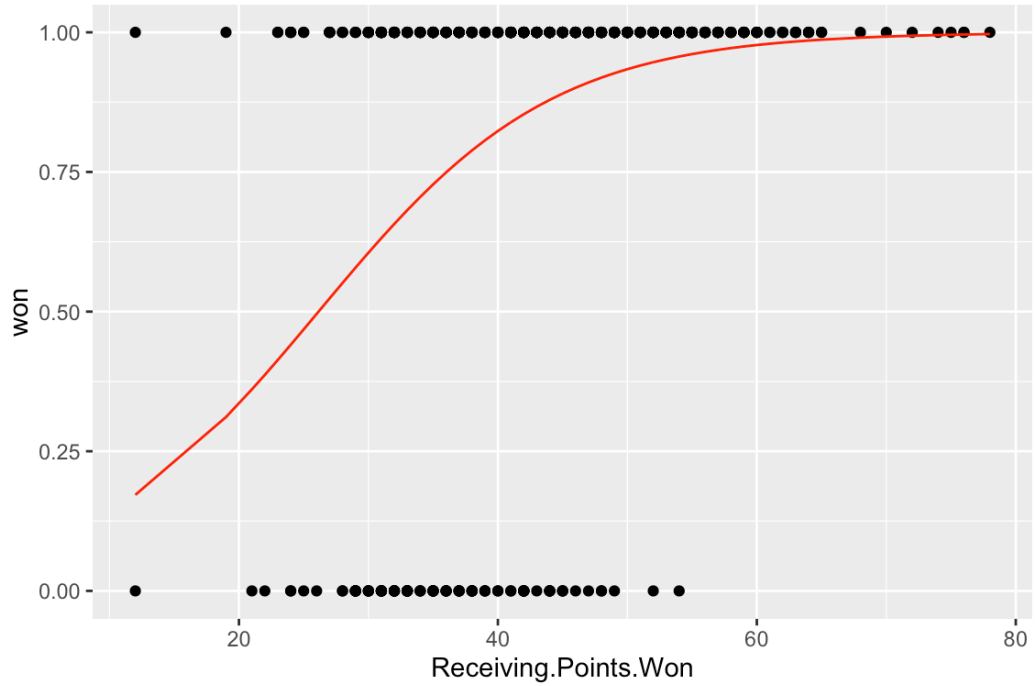
Your turn

The response variable is binary. What type of GLM should be fit?

Model

```
tennis_glm <- glm(won~Receiving.Points.Won, data=tennis,
                  family=binomial(link='logit'))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.91	0.59	-5.0	0
Receiving.Points.Won	0.11	0.02	7.3	0



16/19

Your turn

Write down the fitted model

Resources

- Beginners guide
- Introduction to GLMs
- Quick-R GLMs
- The Analysis Factor, Generalized Linear Models Parts 1-4
- wikipedia
- Do Smashes Win Matches?

Share and share alike