# *General* and *Generalized* Linear Models: Overview and Applications


Michael Otterstatter


BCCDC Biostats Session

May 10, 2019

# Session overview

- In this session we will discuss
- General linear models (GLMs)
  - What are they?
  - How do they work?
  - What are some applications?
- Generalized linear models (GLIMs)
  - What are they?
  - How do they work?
  - What are some applications?

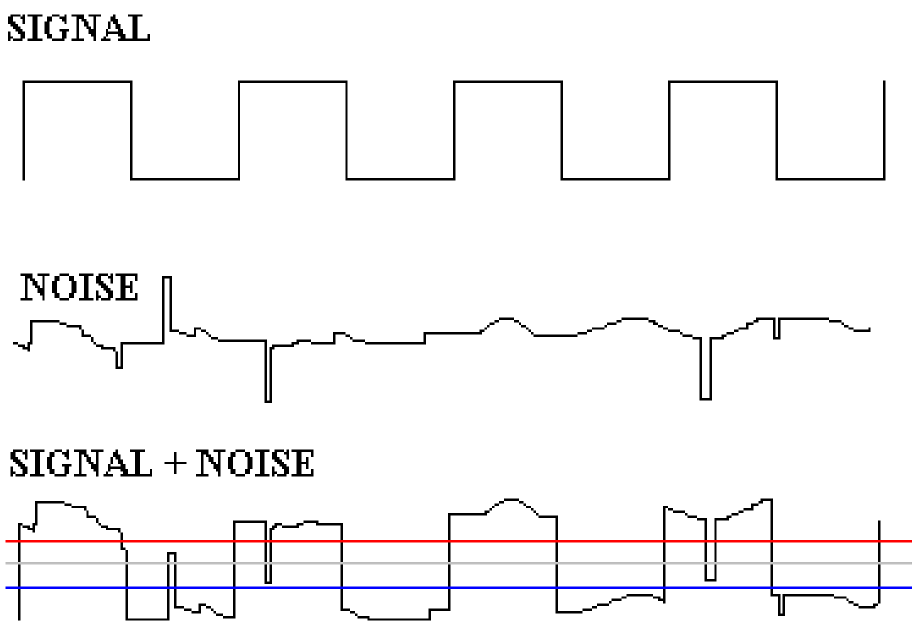
# Background

- When we make observations, we are usually seeing a combination of patterns (‘signal’) and haphazard variation (‘noise’)

# Background

- That is, we think of data as being

***observations = signal + noise***




**Underlying pattern of interest**

**Other sources of variation**

**What we actually observe**

# Background

- Statistical models try to understand the relationship of signal and noise that generates data

***observations = signal + noise***

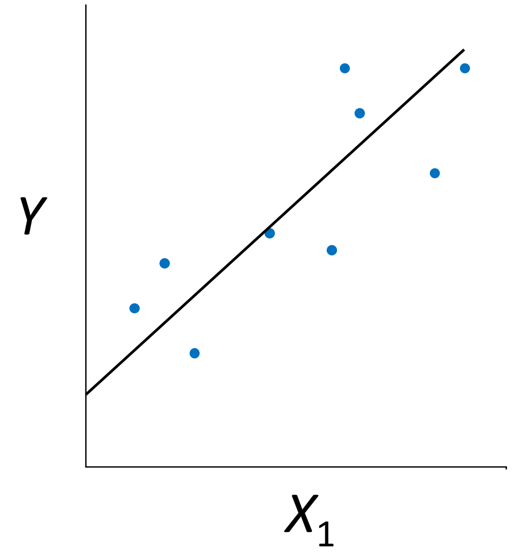
$$\overbrace{Y} = \overbrace{\alpha + \beta_1 X_1} + \overbrace{\varepsilon}$$

- If we understand this relationship, we can
  - succinctly <u>summarize</u> patterns
  - <u>explain</u> observed patterns
  - <u>predict</u> future patterns

# Background

Statistical models often use an equation of a straight line to represent observations as pattern + noise

*"The line of best fit"*




$$Y = \alpha + \beta_1 X_1 + \varepsilon$$

*"The equation of a straight line"*

# A bit of history

- Classical linear models, like linear regression, were first used to study position of astronomical bodies by Gauss in the early 1800s
- Variability in such measurements was largely due to measurement error, for which the Normal or Gaussian distribution was used
- Even at that time, Gauss showed that linear models are sensitive to particular aspects of the data: equal variance, independence of observations, and normally distributed errors

# Linear model basics

The **general linear model (GLM)** represents the pattern with a set of *parameters* (e.g., $\alpha$, $\beta$) plus unexplained ('residual' or 'error') variation ($\varepsilon$)

intercept parameter

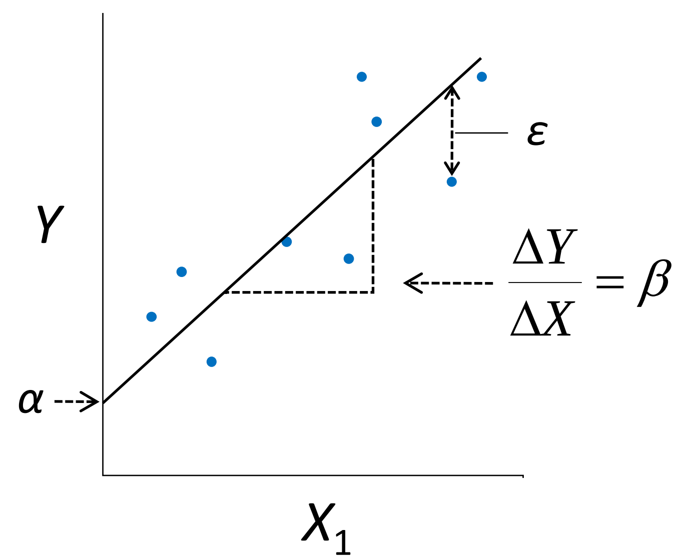
predictor variable

$$Y = \alpha + \beta_1 X_1 + \varepsilon$$

response variable

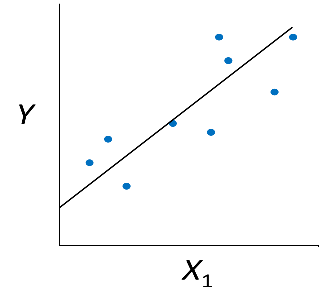
slope parameter

error 'residual'

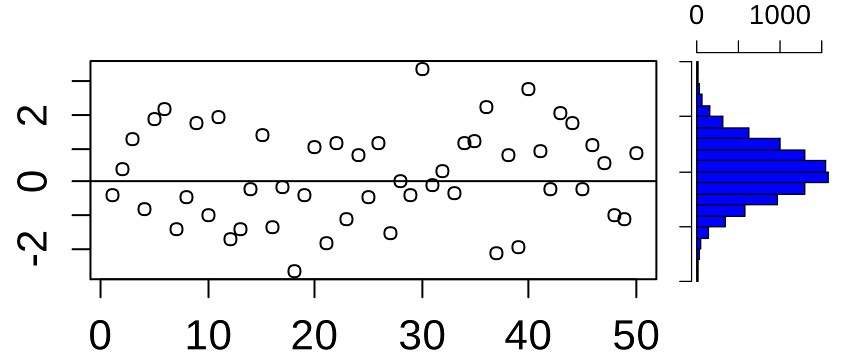


# Linear model assumptions

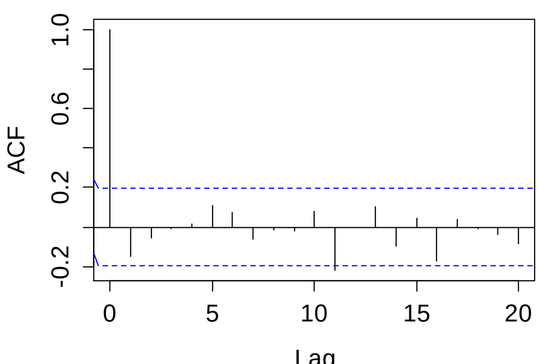
- Relationship between response and predictor(s) is **linear**




- Errors are **normally distributed** and have **constant variance**




- Errors are **independent** of one another



# What is linear in the model?

- A true linear model is linear in its *parameters*. Hence, a linear regression is based on an equation that is linear in its parameters *but may not be linear in its variables*

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

Polynomial regression *is* a linear model

# What is linear in the model?

Which of these are linear models?

A. $log(\mu/t) = \alpha + \beta x$

B. $\mathrm{logit}(\pi_i) = \log \dfrac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 X_i$

C. $Y = \beta_0 + \beta_1 X + \beta_2 X^2$

D. $y_i = \beta_0 + (0.4 - \beta_0)e^{-\beta_1(x_i - 5)}$

# What is linear in the model?

- Keep in mind, lots of non-linear models can be made linear in their parameters by transformations:

$$Q = AP^{\beta}e^{u}$$

$$\ln Q = \ln A + \beta \ln P + u$$

$$q = \alpha + \beta p + u$$

with $q = \ln Q, p \;\; \ln P$, and $\alpha = \ln A$.

# A cautionary tale

## "Anscombe's quartet"

| Dataset 1 | | Dataset 2 | | Dataset 3 | | Dataset 4 | |
|---|---|---|---|---|---|---|---|
| **x** | **y** | **x** | **y** | **x** | **y** | **x** | **y** |
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.1 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.1 | 4 | 5.39 | 19 | 12.5 |
| 12 | 10.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

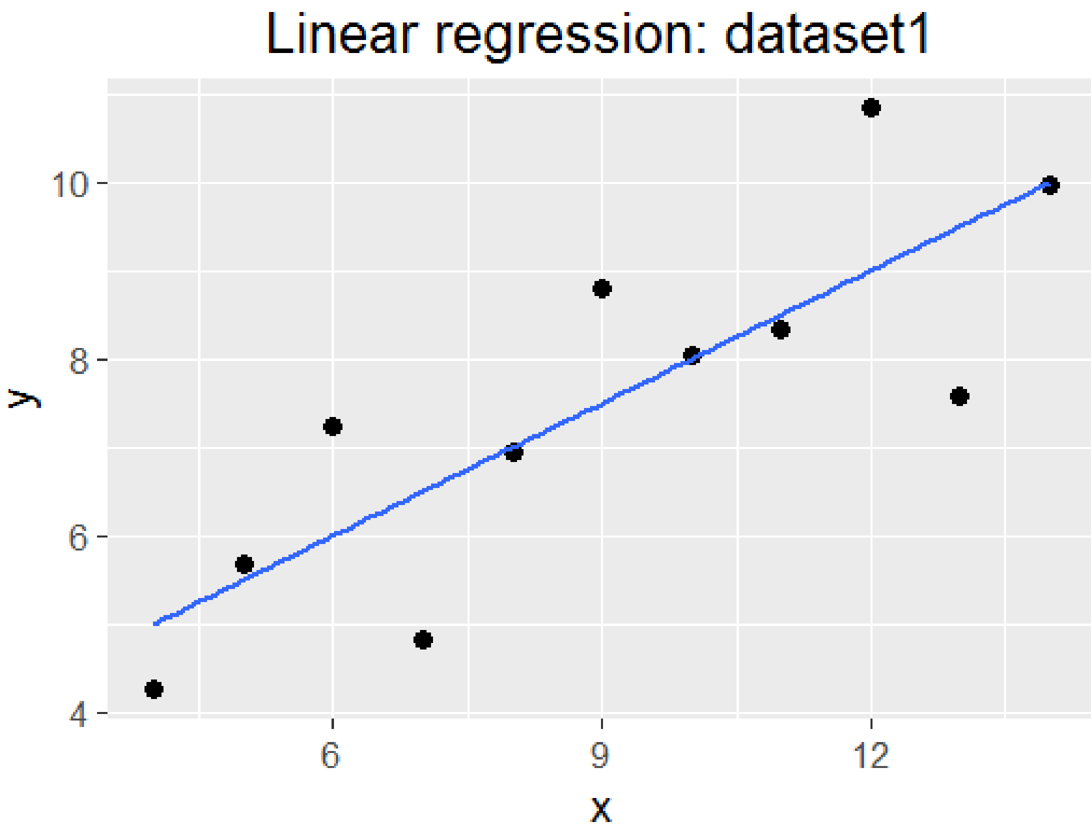
Anscombe, F. J. 1973. "Graphs in Statistical Analysis". American Statistician. 27: 17–21.

# A cautionary tale

In each of the four datasets,

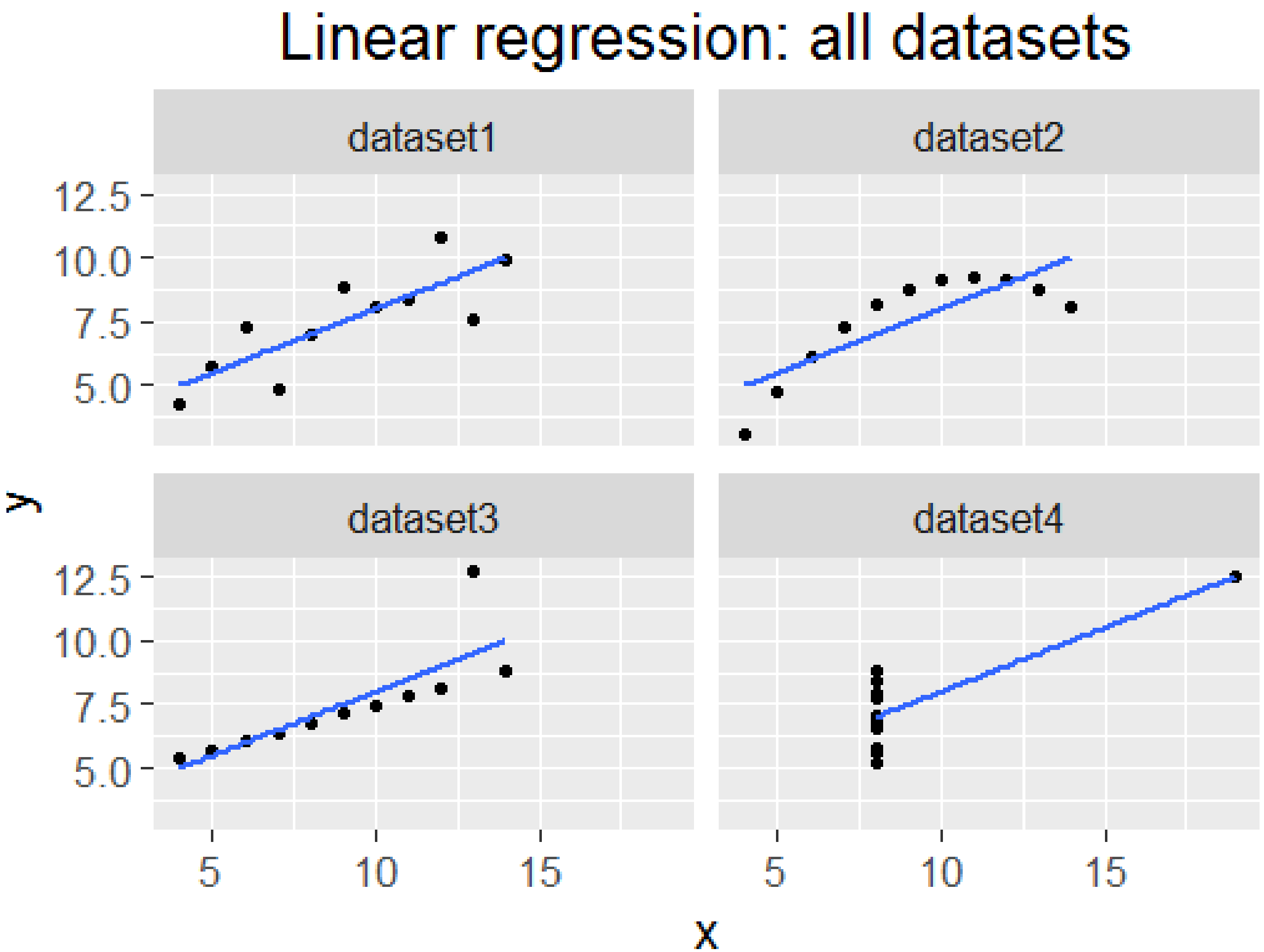
- Mean of x: 9.00
- Variance of x: 11.00
- Mean of y: 7.50
- Variance of y: 4.125
- Correlation between x and y: 0.816

# A cautionary tale




| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| Intercept | 3.000 | 1.125 | 2.667 | 0.026 |
| x | 0.500 | 0.118 | 4.241 | 0.002 |

# A cautionary tale




Regression line is
*y = 3.00 + 0.500x*
for all four datasets!

# A cautionary tale

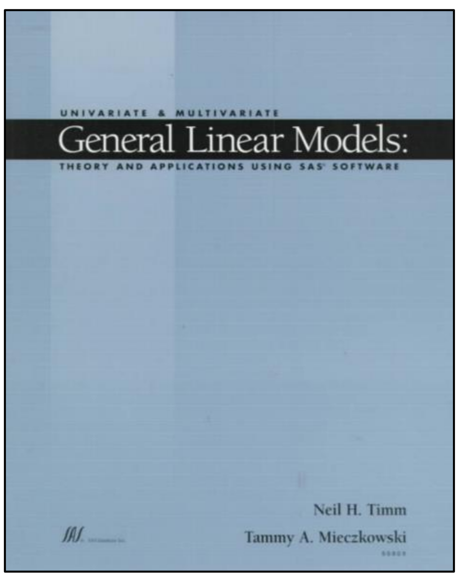
- Lessons from Anscombe's quartet
  - Specifying an appropriate regression model requires careful examination of the data
  - Linear regression is limited for capturing patterns and can imply relations that do not exist
  - Healthy skepticism toward linear regression results is warranted given how easily things can go wrong (also see Chatterjee & Firat 2007. Am Stat 61:248)

# Applications of linear models

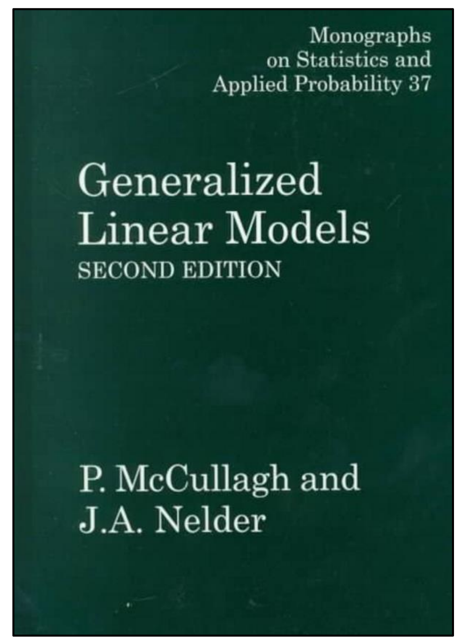
- General linear models include
  - Linear regression (most common form)
  - ANOVA (special case of linear regression)
  - ANCOVA
  - MANOVA (multivariate form of ANOVA)
  - MANCOVA (multivariate form of ANCOVA)
  - t-test
  - F-test

# What about generalized linear models?

- What's the difference between a *general linear model* and a *generalized linear model?*




?



# General vs generalized

| | General linear model | Generalized linear model |
|---|---|---|
| **Typical estimation method** | Least squares, best linear unbiased prediction | Maximum likelihood or Bayesian |
| **Examples** | ANOVA, ANCOVA, linear regression | linear regression, logistic regression, Poisson regression, gamma regression, general linear model |
| **Extensions and related methods** | MANOVA, MANCOVA, linear mixed model | generalized linear mixed model (GLMM), generalized estimating equations (GEE) |
| **R package and function** | lm() in stats package (base R) | glm() in stats package (base R) |
| **Matlab function** | mvregress() | glmfit() |
| **SAS procedures** | PROC GLM, PROC REG | PROC GENMOD, PROC LOGISTIC (for binary & ordered or unordered categorical outcomes) |
| **Stata command** | regress | glm |
| **SPSS command** | regression, glm | genlin, logistic |

https://en.wikipedia.org/wiki/Comparison_of_general_and_generalized_linear_models

# Generalized linear models (GLIMs)

- GLIMs expand on general linear models in two important ways:
  1. Response variable *Y* assumed to have a distribution from the **exponential family**
     - Normal (ordinary linear regression, ANOVA, etc.):
     - Gamma
     - Binomial (logistic regression)
     - Negative binomial
     - Multinomial
     - Poisson
     - and others (inverse Gaussian, negative binomial, zero-inflated Poisson and negative binomial)

# Generalized linear models (GLIMs)

- GLIMs expand on general linear models in two important ways:
  2. The expected value of the response variable ($\mu_i$) is related to a linear equation of predictors through a **link function** ($g$)

$$g(\mu_i) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

  - Note that in ordinary linear models (Normal distribution) the link function is simply the 'identity' of $\mu_i$:

$$\mu_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

# Generalized linear models (GLIMs)

- In Binomial (logistic) regression, a **logit link** is typically used

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

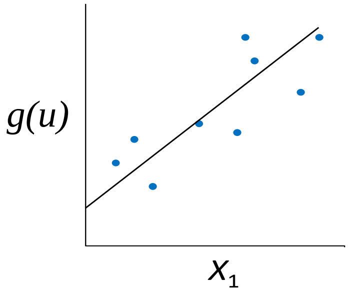
- In Poisson regression, a **log link*** is typically used

$$\log(\mu_i) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

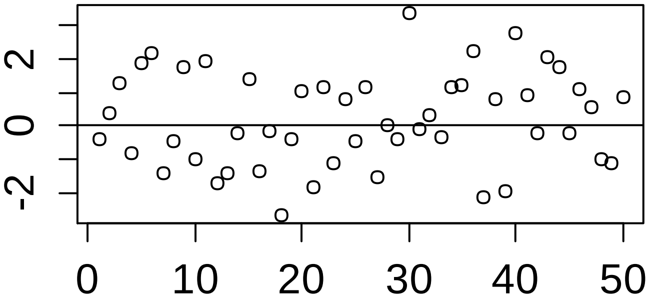
* natural (base *e*) logarithm

# Assumptions of GLIMs

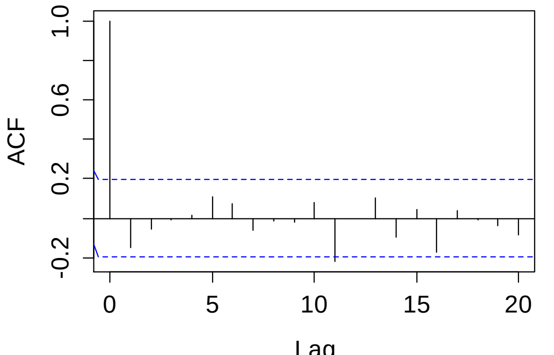
- Relationship between *transformed response* $g(\mu_i)$ and predictor(s) is **linear**




- *Residuals* should be evenly distributed and have **constant variance** (but not assumed to be normally distributed)

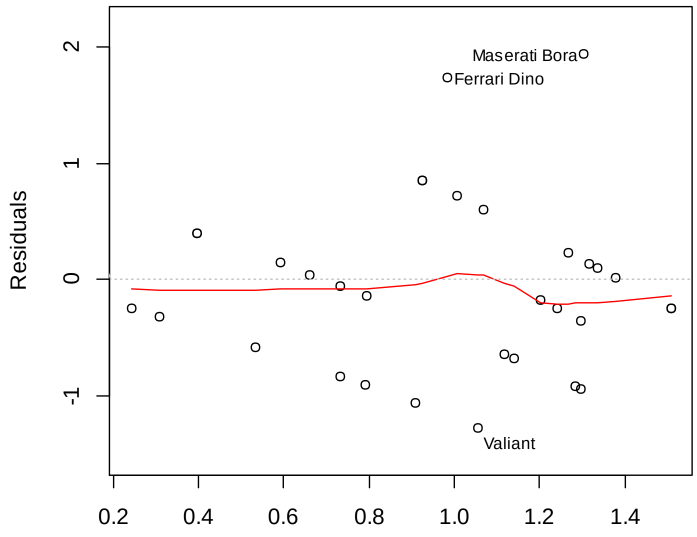


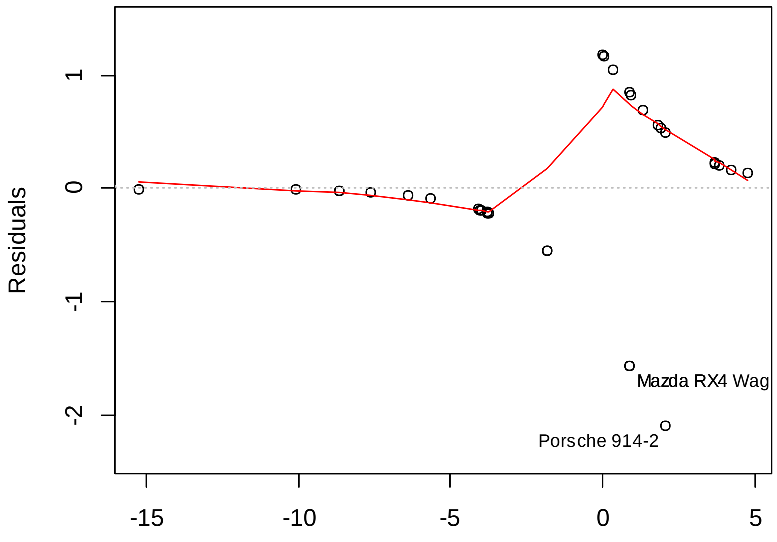

- Residuals are **independent** of one another



# An aside

- Unlike general linear models, GLIM residuals are not always very helpful for model diagnostics
  - In binomial models, residuals can have only two possible values for a given model fit
  - In small datasets, binomial and Poisson residuals tend to show curved lines of points



# General vs generalized

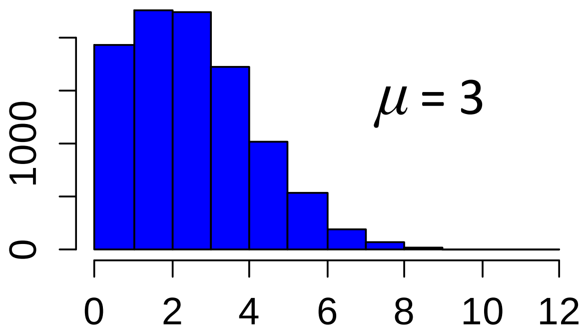
- For **general linear models** (e.g., ordinary linear regression), parameter estimates are calculated using least squares
  - output includes family measures such as F-tests, sums-of-squares, R-squared, etc.

- However, for **GLIMs**, estimation is typically done using maximum likelihood, which finds solutions via numerical optimization
  - output includes less family measures such as deviance, AIC, dispersion, likelihood ratio chi-square, etc.)

# Poisson regression as an example

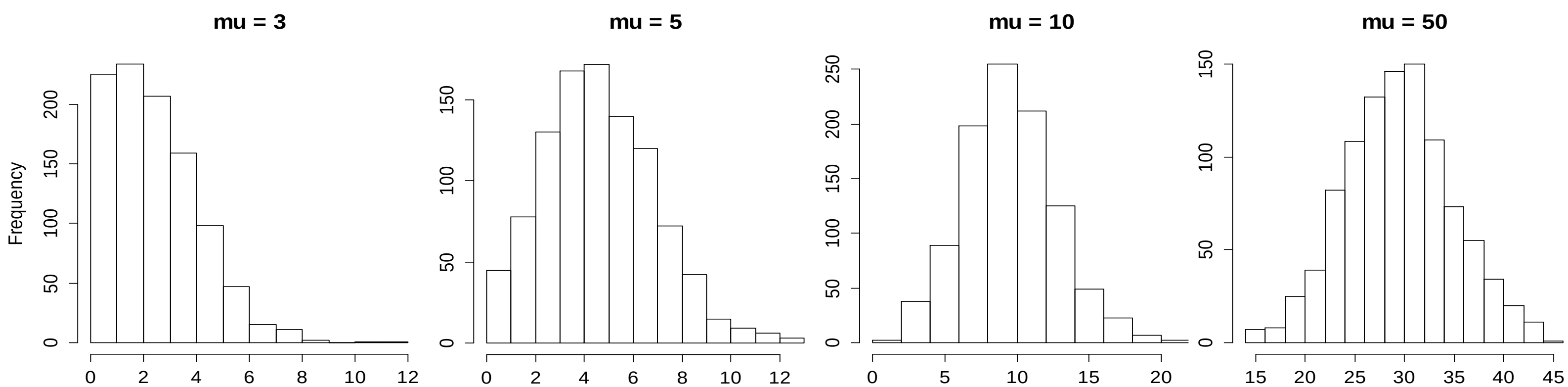
- **Poisson regression** assumes the underlying data generating process produces rare, random events (discrete, non-negative counts)
  - where 'rare' is meant relative to the large number of events that could possibly have occurred (but didn't) in the sampling unit or interval

- Examples:
  - car crashes on a particular stretch of road
  - phone calls to a switchboard
  - Particle emissions due to radioactive decay



# An aside

- Poisson distributions with small mean values (μ < 10) are distinctly skewed, but with larger means (μ >10), they increasingly approximate the Normal distribution




- Yet, Poisson regression is distinctly different from ordinary linear regression (non-negative integers only, multiplicative effects, etc.)

# Poisson regression as an example

- If we write out the formal definition of a Poisson model, it helps to clarify what assumptions are being made:

$$y_i \sim Poisson(\mu_i)$$
$$\log(\mu_i) = \alpha + \beta_1 x_1 + \beta_2 x_2 \ldots$$

  - the observed count $y$ in sampling unit $i$ is assumed to come from a Poisson distribution with mean $u$ specific to that unit $i$ (see Appendix slides for further explanation)
  - the natural log of $u_i$ is assumed to be a linear function of the regression variables $x_1$, $x_2$…

# Poisson regression as an example

- Modeling $y$ as observations from a Poisson distribution carries different assumptions than ordinary linear regression
  - In Poisson regression we assume variance of our observations is proportional to the mean (hence, no separate parameter for variance, as in the Normal distribution)

$$y \sim Poisson(\mu)$$

$$y \sim Normal(\mu, \sigma^2)$$

  - In fact in all GLIMs the variance of the response $y$ is related to the mean $\mu$ through a variance function $V$ with ***dispersion parameter*** $\phi$

$$var(y_i) = \frac{\phi V(\mu_i)}{w_i}$$

# Poisson regression as an example

- Whenever running a Poisson (or binomial) regression, we need to check this assumption of variance proportional to the mean, i.e., that the dispersion parameter $\phi = 1$

| Criteria For Assessing Goodness Of Fit | | | |
|---|---|---|---|
| Criterion | DF | Value | Value/DF |
| Deviance | 2 | 2.8207 | 1.4103 |
| Scaled Deviance | 2 | 2.8207 | 1.4103 |
| Pearson Chi-Square | 2 | 2.8416 | 1.4208 |
| Scaled Pearson X2 | 2 | 2.8416 | 1.4208 |
| Log Likelihood | | 837.4533 | |
| Full Log Likelihood | | -16.4638 | |
| AIC (smaller is better) | | 40.9276 | |
| AICC (smaller is better) | | 80.9276 | |
| BIC (smaller is better) | | 40.0946 | |

Measures of deviance divided by the model degrees of freedom should be roughly equal to 1.0

If not, variance is likely greater than (***overdispersion***) or less than (***underdispersion***) proportional to the mean

# Poisson regression as an example

- Poisson regression usually involves some notion of time (e.g., intervals between events, 'inter-arrival times') or 'exposure' and therefore naturally suited to modeling **rates**

$$\log(\mu_i) = \log(n_i) + \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

  - Where log(*n*) is the natural log of the exposure variable, aka the '**offset'**, e.g., time, population size, or any other denominator for the rate

  - Equivalently: $\log(\frac{\mu_i}{n_i}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots$

# Poisson regression as an example

- Running Poisson models is just as easy as ordinary linear regression!
  - In SAS:

```
proc genmod;
    class outcome treatment;
    model counts = outcome treatment / dist=poisson link=log offset=ln;
run;
```

  - In R:

```
poisson.fit <- glm(counts ~ outcome + treatment, family = poisson())
```

# Applications of GLIMs

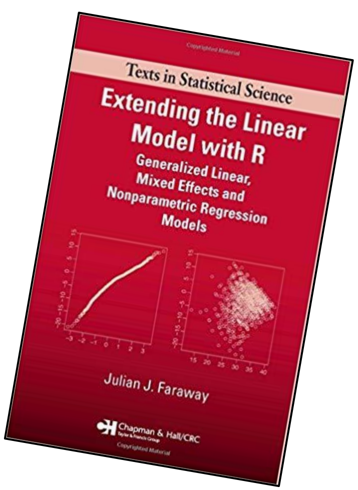
- Generalized linear models include
  - Linear regression, ANOVA (continuous outcome)
  - Binomial regression (binary outcome, probabilities)
  - Poisson regression (counts, rates)
  - Multinomial regression (multiple categorical outcome)
  - Gamma regression (flexible; could be used for survival data)

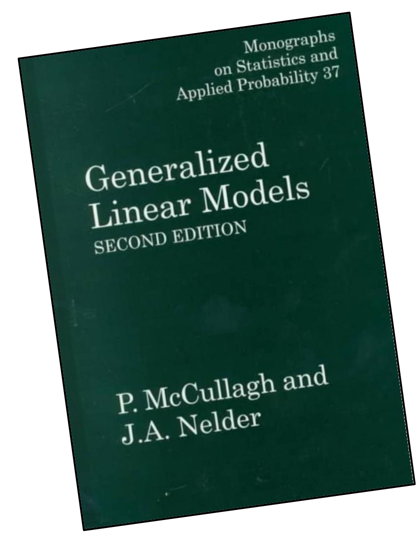
# Summary

- Generalized linear models (GLIMs) are fundamentally similar to ordinary linear models, but with enhancements that allow analysis of many types of data (counts, binary outcomes, categorical data)
- GLIMs make certain assumptions, similar to general linear models, that must be assessed to ensure model validity
- Certain technical aspects of GLIMs sound confusing (deviance, dispersion, maximum likelihood) but generally have easily understood meanings – ask your favourite statistician ☺

# Useful references

- McCullagh P, Nelder JA. 1989. *Generalized Linear Models*. Chapman & Hall.
- Faraway, J. J. (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.



# Appendix: clarifying points

Imagine a data set with $i$ observed counts that we wish to use in a regression, and each observed count $y_i$ has an associated covariate value $x_i$. If we fit a Poisson regression, we are assuming that each count $y_i$ comes from a Poisson distribution with a mean $u_i$

$y_i$ ~ Poisson($u_i$)

This is the same thing as saying the expected value of $y_i$ is $u_i$ (sometimes written as E($y_i$) = $u_i$).

The 'expected value' refers to the fact that each observation $y_i$ is a draw from a Poisson distribution and therefore could have been any of a range of values, but the most likely value (or the mean value, if we took many draws from this Poisson distribution) is $u_i$.

Observations are assumed to be independent of one another, so <u>each $y_i$ has an expected value $u_i$</u>. Remember that although a Poisson distribution has only a single parameter $u$, our regression is not just referring to one Poisson distribution but $i$ Poisson distributions (one for each observation $y$).

The left-hand side of the regression equation

$g(\mu_i)=\alpha+\beta x_i$

is therefore not fixed, but has $i$ values. The Poisson regression is attempting to estimate what values of $\alpha$ and $\beta$ provide the best fit to these expected values, given the covariate values $x_i$.

When the model calculates the best-fit (maximum likelihood) estimates of $\alpha$ and $\beta$, we can plug these into the equation above to get the predicted values for each observation.

# Appendix: clarifying points

Although it sounds a bit different, the concept of random variation is the same in generalized linear models as in general linear models.

In ordinary linear regression, we model the observed *y*'s as expected values from the equation $y = a + bx$, plus some random error (often written as $y = a + bx + \boldsymbol{e}$). Sometimes for general linear models this is written as the expected value of y, given x

$E(y|x) = a + bx$

In generalized linear models, we do not refer to random error in the model equation; instead, the random error around our observations is assumed to come from the underlying distribution, for example

$y \sim Poisson(u)$

So we still have the same concept of random variation, but it is shown differently in generalized linear models and assumed to follow different distributions: in general linear models, random error is assumed to be Normal; in generalized linear models, the random error can come from any of the exponential distributions: Poisson, binomial, gamma, etc.

# Appendix: clarifying points

To fit a generalized linear model, imagine a process whereby an algorithm finds the straight line that best fits the observations $y_i$. The parameters of this straight line are our maximum likelihood estimates for $\alpha$ and $\beta$. If we plug these estimates into the model equation

$g(\mu_i)=\alpha+\beta x_i$

we will get the expected value $\mu_i$ for each observation $y_i$.

Of course, the observations $y_i$ are not identical to the expected values $\mu_i$; instead, there is some additional variation that we call the residual, or 'error'.

Just as in general linear models, generalized linear models have specific expectations about this residual variation, e.g., in Poisson and Binomial models, the variation should be proportional to the mean.