Generalized Linear Models
(Regression: Linear, Logistic, Poisson)

Nagiza F. Samatova, samatova@csc.ncsu.edu
Professor, Department of Computer Science
North Carolina State University

Senior Scientist, Computer Science & Mathematics Division
Oak Ridge National Laboratory

Supervised Learning: High-level Problem Stmt

- **Known** for each observation in **the sample of size n** from the population**:**
  - **Predictor / explanatory** variables: $X_1, X_2, \dots, X_q$
    - **Continuous** and/or
    - **Categorical**
  - **Response / outcome** variable: Y
    - **Continuous** or
    - **Categorical**
- **Unknown** for the entire population:
  - Relationship (**model**) between the response variable Y and a set of ***q predictor variables*** $\vec{X} = (X_1, X_2, \dots, X_q)$ and/or their ***transformations***
  - A set of **m parameters** of this model: $\vec{\beta} = (\beta_0, \beta_1, \dots, \beta_m)$

$$\boxed{f\colon \vec{X} \to Y \text{ such that } Y = f(\vec{X}, \vec{\beta})}$$

Example

Predictor Variables Response

Living Area	# of rooms	Distance	Rent
230	1	2.1	600
506	2	3.2	1000
433	2	1.0	1100
109	1	1	500
---	…		
150	1	1.5	?
270	1.5	3.0	?

Univariate vs. Multivariate Model

Univariate ($q = 1$): The number of predictor variables is one

Y = Rent
X = Living Area

Y = Rent
X = Distance

Multivariate ($q > 1$): The number of predictor variables is more than one

Y = Rent
X_1 = Area
X_2 = Rooms
X_3 = Distance

Living Area	# of rooms	Distance	Rent
230	1	2.1	600
506	2	3.2	1000
433	2	1.0	1100
109	1	1	500
---	…		
150	1	1.5	?
270	1.5	3.0	?

Known: Predictors & their Transformations

Unknown: Parameters

$$\text{Rent} = \beta_0 + \beta_1 * Area^2 + \beta_2 * Rooms * exp\{Distance\}$$

Unknowns in a Linear Model:
Conditional Mean of the Response & Weights

We want to estimate the **conditional mean of the response** as the ***weighted linear*** combination (**sum, Σ**) of the predictor variables $X_1, X_2, \ldots, X_q$ such that:

$$\boldsymbol{Y = \mu_Y + error}$$

$\boldsymbol{\mu_Y}$ **is linear in parameters / weights** $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_m}$

We want to predict the mean of the Y distribution for observations with a given set of predictor variables by applying the proper weights $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_m}$ on the **predictors** $X_1, X_2, \ldots, X_q$ and/or **their transformations.**

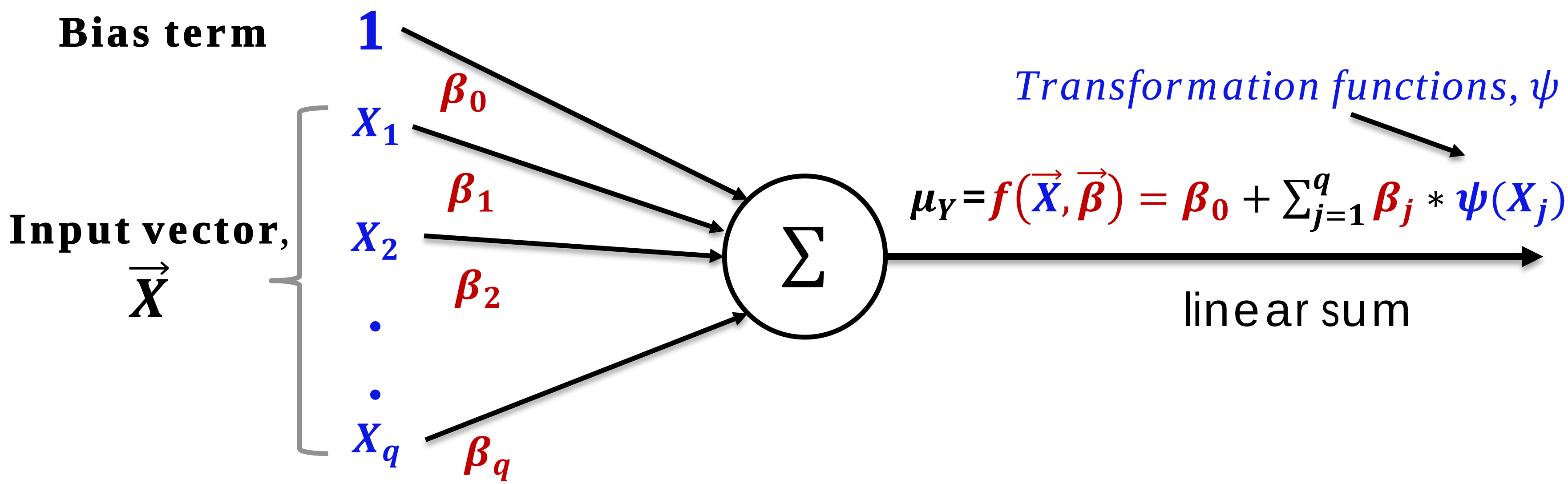
$\boldsymbol{f\colon \vec{X} \to \mu_Y \ :\ \mu_Y = f(\vec{X}, \vec{\beta}) \ and\ f(\vec{X}, \vec{\beta})}$ **is linear in** $\boldsymbol{\vec{\beta}}$

Linear Models: Homogeneous Transformations

$Y = \mu_Y + error$

μ_Y **is linear in parameters / weights** $\beta_0, \beta_1, \dots, \beta_q$

$f: \vec{X} \to \mu_Y \;:\; \mu_Y = f(\vec{X}, \vec{\beta})$ *and* $f(\vec{X}, \vec{\beta})$ **is linear in** $\vec{\beta}$



Ex: $\mu_Y = f(\vec{X}, \vec{\beta}) = \beta_0 + \sum_{j=1}^{q} \beta_j * X_j$

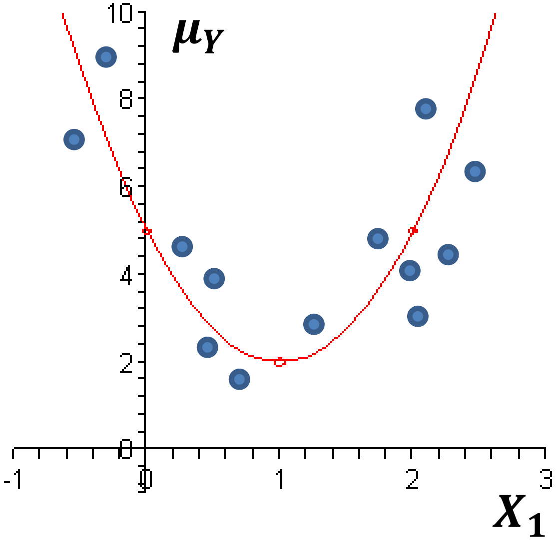
Ex: $\mu_Y = f(\vec{X}, \vec{\beta}) = \beta_0 + \sum_{j=1}^{q} \beta_j * log(X_j)$

Ex: $\mu_Y = f(\vec{X}, \vec{\beta}) = \beta_0 + \sum_{j=1}^{q} \beta_j * sin(X_j)$

Ex: $\mu_Y = f(\vec{X}, \vec{\beta}) = \beta_0 + \sum_{j=1}^{q} \beta_j * exp(X_j)$

Example: Univariate Linear Regression

$$\boxed{\mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2}$$



Known: $\psi_1(X_1) = X_1$ **and** $\psi_2(X_1) = X_1^2$

Unknown: $\beta_0, \beta_1, \beta_2$

- We are NOT learning the transformation functions; they should be provided as input
- We are learning the weights on those functions

$$\boxed{\mu_Y = f(\vec{X}, \vec{\beta}) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2}$$

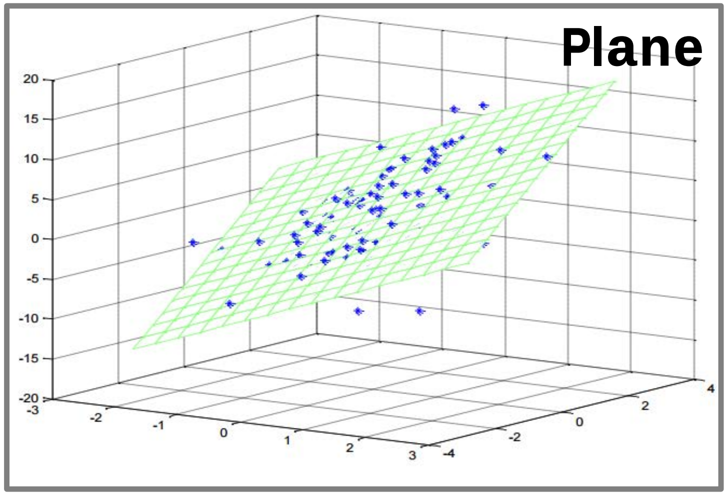
$$\boxed{\frac{\partial f(\vec{X}, \vec{\beta})}{\partial \beta_1} = X_1}$$

$$\boxed{\frac{\partial f(\vec{X}, \vec{\beta})}{\partial \beta_2} = X_1^2}$$

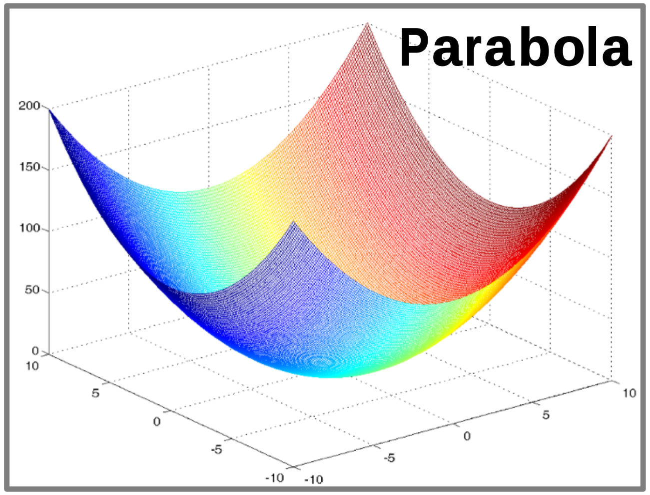
$$\boxed{\frac{\partial f(\vec{X}, \vec{\beta})}{\partial \beta_0} = 1}$$

Examples: Multivariate Linear Regression

Multivariate Linear Regression of Two Predictors



$$\mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$



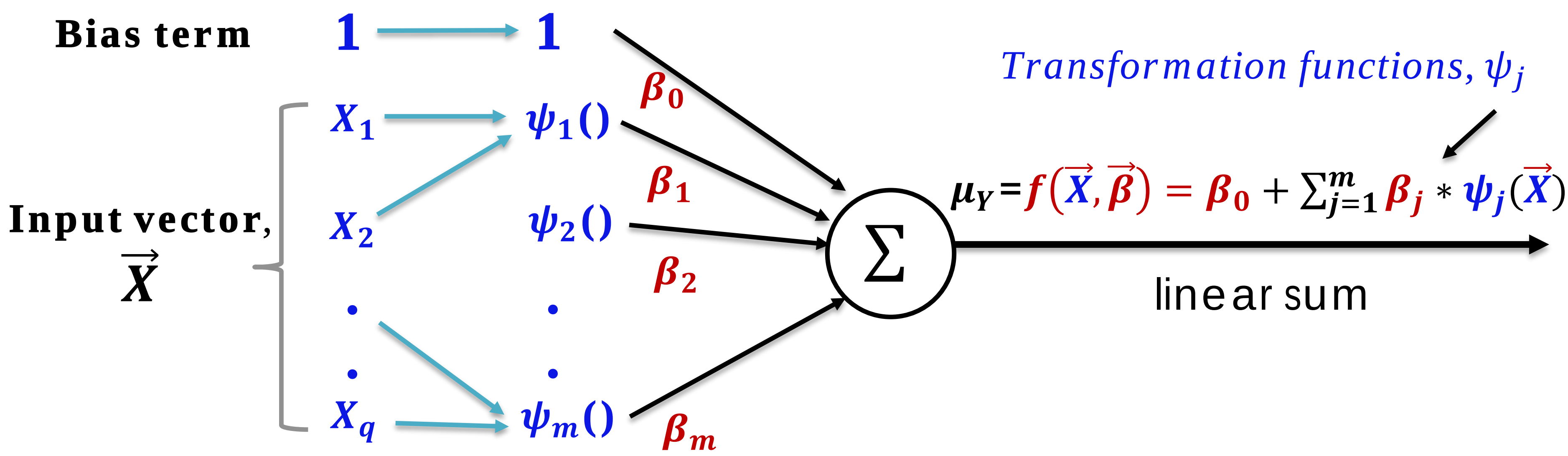
$$\mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \\ + \beta_3 X_2 + \beta_4 X_2^2 + \beta_5 X_1 X_2$$

Linear Models: Heterogeneous Transformations

$Y = \mu_Y + error$

μ_Y is **linear** in **parameters / weights** $\beta_0, \beta_1, \ldots, \beta_q$

$f: \vec{X} \to \mu_Y \;:\; \mu_Y = f(\vec{X}, \vec{\beta})$ *and* $f(\vec{X}, \vec{\beta})$ is **linear** in $\vec{\beta}$



$$\mu_{Rent} = \beta_0 + \beta_1 * Area^2 + \beta_2 * Rooms * exp\{Distance\}$$

$$\psi_1(\vec{X}) = \psi_1(X_1) = Area^2$$

$$\psi_2(\vec{X}) = \psi_2(X_2, X_3) = $$
$$= Rooms * exp\{Distance\}$$

Generalized Linear Model:
Link Function & Weights

We want to estimate **the function of the conditional mean of the response (the link function)** as the ***weighted linear*** combination of the predictor variables $X_1, X_2, \dots, X_q$ such that:

$$\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{\mu_Y}) + \boldsymbol{error}$$

$\boxed{\boldsymbol{g}(\boldsymbol{\mu_Y}) \textbf{ is linear in parameters / weights } \boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \dots, \boldsymbol{\beta_m}}$

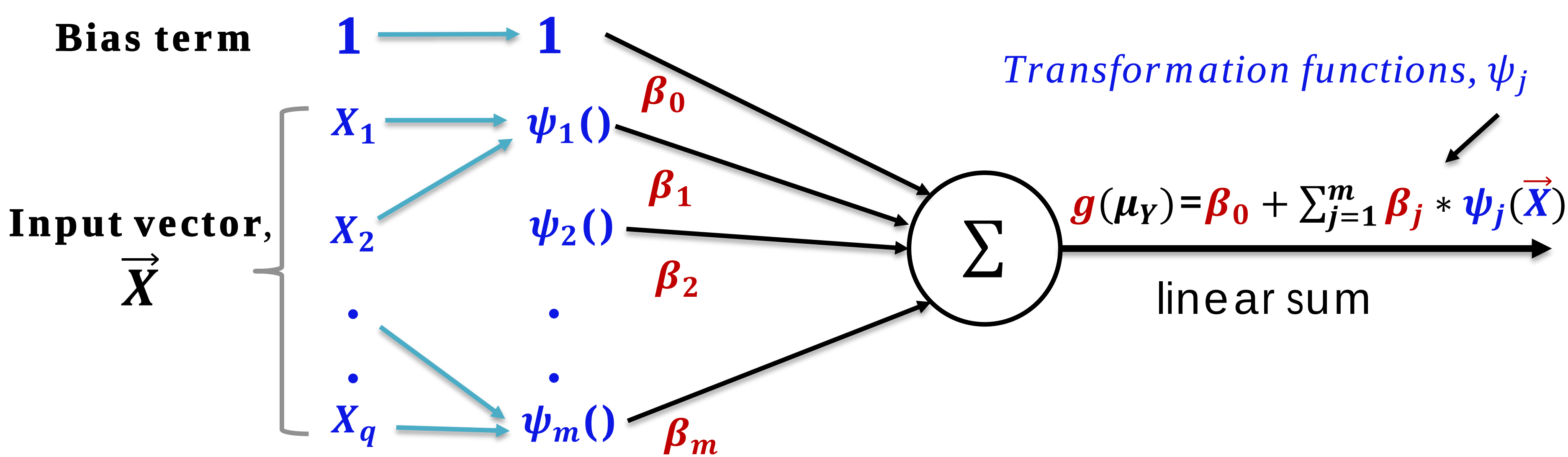
We want to predict the function $g(\mu_Y)$ of the mean of the Y distribution for observations with a given set of predictor variables by applying the proper weights $\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \dots, \boldsymbol{\beta_m}$ on the **predictors** $X_1, X_2, \dots, X_q$ and/or **their transformations**.

Generalized Linear Models: Schematics

$Y = g(\mu_Y) + error$

$\boxed{g(\mu_Y) \text{ is linear in parameters } \beta_0, \beta_1, \dots, \beta_q}$

$$\boxed{f: \vec{X} \to \mu_Y \ : \ g(\mu_Y) = f(\vec{X}, \vec{\beta}) \text{ and } f(\vec{X}, \vec{\beta}) \text{ is linear in } \vec{\beta}}$$



$$\boxed{\log(\mu_{Rent}) = \beta_0 + \beta_1 * Area^2 + \beta_2 * Rooms * \exp\{Distance\}}$$

$\psi_1(\vec{X}) = \psi_1(X_1) = Area^2$

$\psi_2(\vec{X}) = \psi_2(X_2, X_3) =$
$= Rooms * \exp\{Distance\}$

Linear vs. Nonlinear Model

Response = Linear Combination of Explanatory Variables

Function of the Response = Linear Combination of Explanatory Variables

linear model
in terms of $\beta's$,
unknown parameters
(**glm()** in **R**)

$$\left\{\begin{array}{l} \mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q \\ \mu_Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \underline{\beta_3 X_1^3} \\ g(\mu_Y) = \beta_0 + \beta_1 X_1 + \underline{\beta_2 X_1 X_2} \quad \text{coupled predictors} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{polynomial predictors} \\ g(\mu_Y) = \beta_0 + \beta_1 X_1 + \underline{\beta_2 \exp^{X_1}} \quad \text{transformed predictors} \\ g(\mu_Y) = \beta_0 + \underline{\beta_1 log X_1} + \underline{\beta_2 \sin(X_2)} \end{array}\right.$$

Known: $X_1, X_2, \ldots,$
Unknown: $\beta_0, \beta_1, \ldots$

$$\boxed{\text{The equation is linear in the parameters } (\beta_0, \beta_1, \ldots, \beta_q)}$$

***non*-linear model**
in terms of $\beta's$, unknown parameters (**nls() in R**)

$$\mu_Y = \beta_0 + \beta_1 exp^{\frac{X}{\beta_2}}$$

GLM Requirements: Predictors & Response

Predictors

- No distributional assumptions about the predictor variables, $X_1, X_2, \ldots, X_q$
- Predictors could be categorical or continuous
- Nonlinear functions of predictors are allowed: X_1^2, $\exp^{X_1}$, $\sin(X_2)$

Response

- Y is normally distributed (i.e., **Gaussian distribution**) <u>or</u>
- Y follows a distribution that is a member of **the exponential family:**
  - **Binomial distribution**
  - **Poisson distribution**

Generalized vs. Standard Linear Model

Generalized Linear Models extend standard linear model (aka linear regression):

- by fitting the ***function of the conditional mean response*** $(g(\mu_Y))$ rather than conditional mean response (μ_Y)

- by assuming that the distribution of the response variable Y is a member of the ***exponential family of distributions*** rather than limited to the normal/Gaussian distribution

- by modeling the response variable that has ***categorical*** outcomes or discrete ***counts*** not only the continuous values

- by estimating the unknown parameters $(\boldsymbol{\beta_0}, \boldsymbol{\beta_1}, \ldots, \boldsymbol{\beta_m})$ via ***Maximum Likelihood Estimation*** (**MLE**) rather than ***Ordinary Least Squares*** (**OLS**)

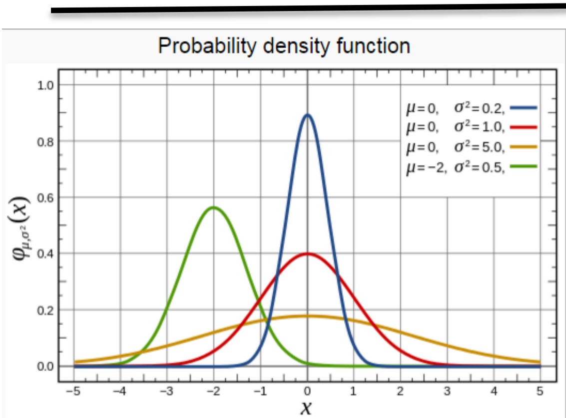
Response: Continuous, Binary, Polytomous, Count

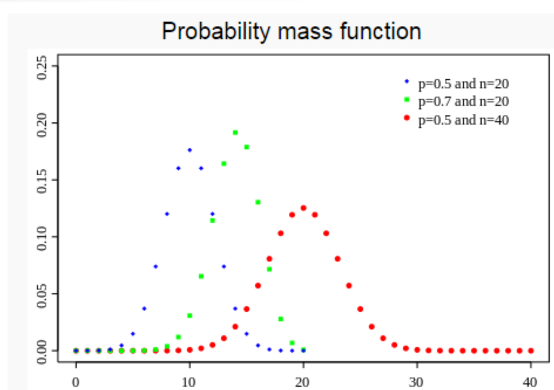
RESPONSE, Y

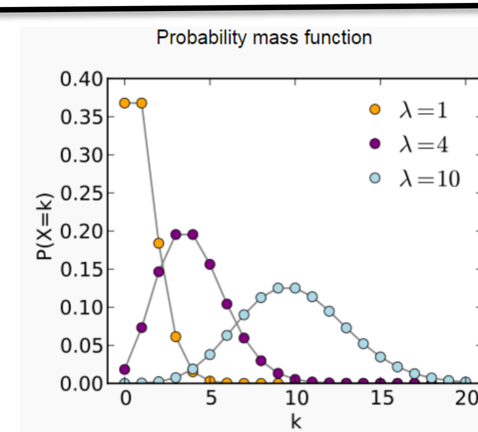
Continuous	Categorical	Count
Temperature Humidity Stock price	**Binary** Yes/No Passed/Failed **Polytomous: Ordinal** Poor/good/excellent **Polytomous: Nominal** Republican/democrat/independent	Number of traffic accidents in a week Number of coffee cups per day

Response, Y: Distribution $f(Y|mean, var)$

Gaussian	Binomial	Poisson
	distribution of the # of successes in a sequence of *n* independent yes/no experiments, each of which yields success with probability $p = P(Y=1)$	the probability of a given # of events occurring in a fixed interval of time/space if these events occur with a known average rate, independently of the time since the last event.
Mean: μ_Y	Mean: $\mu_Y = np$	Mean: $\mu_Y = \lambda$
Variance: σ^2	Variance: $np(1-p)$	Variance: λ
$\frac{1}{\sigma\sqrt{2\pi}} exp^{\frac{-(Y-\mu)^2}{2\sigma^2}}$	$P(Y=k) = \binom{n}{k} p^k (1-p)^{n-k}$	$P(Y=k) = \frac{\lambda^k exp^{-\lambda}}{k!}$







Exponential Family & Link Function

glm (*formula*, family = *family* (link = *function*), data =)

Family	Link Function Name	Link Function: $g(\mu_Y)$
binomial	(link = "logit")	$g(\mu_Y) = log_e\left(\frac{\mu_Y}{1-\mu_Y}\right) = log_e\left(\frac{p}{1-p}\right)$
gaussian	(link = "identity")	$g(\mu_Y) = \mu_Y$
poisson	(link = "log")	$g(\mu_Y) = log_e\mu_Y = log_e\lambda$

Logistic Regression:

glm (*Y~X1+X2+X3*, family = *binomial* (link = *"logit"*), data =mydata)

Poisson Regression:

glm (*Y~X1+X2+X3*, family = *poisson* (link = *"log"*), data =mydata)

Linear Regression:

glm (*Y~X1+X2+X3*, family = *gaussian* (link = *"identity"*), data =mydata)

Logit Function in Logistic Regression

Family	Link Function Name	Link Function: $g(\mu_Y)$
binomial	(link = "logit")	$g(\mu_Y) = log_e\left(\frac{\mu_Y}{1-\mu_Y}\right) = log_e\left(\frac{p}{1-p}\right)$

Logistic Regression: Predict a binary outcome from a set of continuous and/or categorical predictor variables

$$g(\mu_Y) = logit(p) = log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q$$

μ_Y: the **conditional mean** of Y, i.e., the probability $p = P(Y = 1)$ that $Y = 1$ given a set of X values

$\frac{\mu_Y}{1-\mu_Y}$: the **odds** that $Y = 1$ given a set of X values

$\log \frac{\mu_Y}{1-\mu_Y}$: the **log-odds** or **logit** that $Y = 1$ given a set of X values

Logistic Regression

Family	Link Function Name	Link Function: $g(\mu_Y)$
binomial	(link = "logit")	$g(\mu_Y)=log_e\left(\frac{\mu_Y}{1-\mu_Y}\right)=log_e\left(\frac{p}{1-p}\right)$

Logistic Regression: Predict a **binary outcome** Y from a set of continuous and/or categorical predictor variables

$$g(\mu_Y) = logit(p) = log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q$$

Y: the **binomial distribution**, with the probability $p = P(Y = 1)$ that $Y = 1$ given a set of X values

Y: the **binary outcome: Y=1 or Y=0**

Continuous logit() function

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q \in \mathbb{R} = (-\infty; +\infty)$$

$$logit(p) \in \mathbb{R} = (-\infty; +\infty)$$

Logit vs. Probability

Family	Link Function Name	Link Function: $g(\mu_Y)$
binomial	(link = "logit")	$g(\mu_Y) = log_e\left(\frac{\mu_Y}{1-\mu_Y}\right) = log_e\left(\frac{p}{1-p}\right)$

Logistic Regression:

$$logit(p) = log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q$$

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q} = e^z$$

$$p = \frac{e^z}{1 + e^z}$$

$$p = \frac{1}{1 + e^{[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q)]}}$$

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q \in \mathbb{R} = (-\infty; +\infty)$$

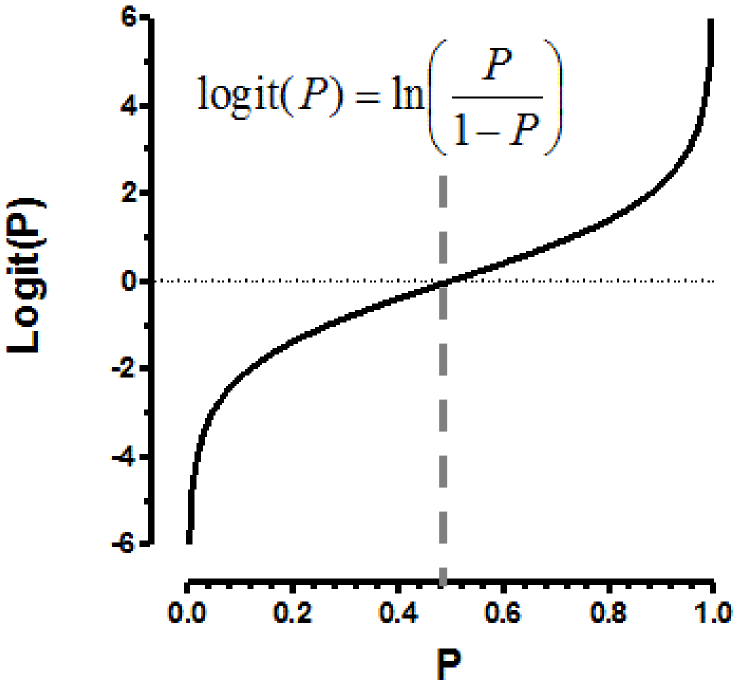
$$logit(p) \in \mathbb{R} = (-\infty; +\infty)$$

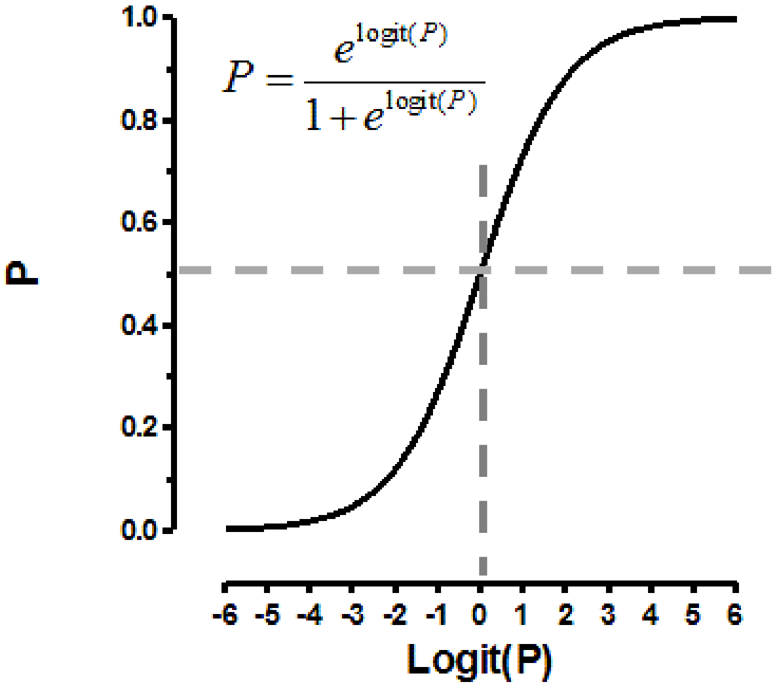
$$p \in (0; 1)$$

Logit vs. Probability

$$logit(p) = log_e\left(\frac{p}{1-p}\right)$$

$$p = \frac{e^{logit(p)}}{1 + e^{logit(p)}}$$





$$logit(p) \in \mathbb{R} = (-\infty; +\infty)$$

$$p \in (0; 1)$$

Example: Logistic Regression: Infidelity

Data Dictionary for the Infidelity Data

affairs	The number of affairs during the past year
gender	
age	
yearsmarried	The number of years in marriage
children	Had children
religiousness	1=anti to 5=very
education	
occupation	7-point classification with reverse numbering
rating	Self-rating of marriage: 1=very happy to 5=very unhappy

```
> t(Affairs[1,])
              4
affairs       "0"
gender        "male"
age           "37"
yearsmarried  "10"
children      "no"
religiousness "3"
education     "18"
occupation    "7"
rating        "4"
```

What personal, demographic, and relationship variables predict marital infidelity?

Outcome, Y: **Binary** (affair/no affair)

```
install.packages("AER")
library(AER)

data(Affairs, package="AER")
summary(Affairs)
names(Affairs)
dim(Affairs)
t(Affairs[1,])
```

Infidelity: Build Logistic Regression Model

```
fit.full <- glm (ynaffair ~
                    gender +
                    age +
                    yearsmarried +
                    children +
                    religiousness +
                    education +
                    occupation +
                    rating,
                  data = Affairs,
                  family = binomial(link="logit")
             )
```

Infidelity: Interpreting the Model

```
> summary(fit.full)

Call:
glm(formula = ynaffair ~ gender + age + yearsmarried + children + 
    religiousness + education + occupation + rating, family = binomial(link = "logit"), 
    data = Affairs)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.5713  -0.7499  -0.5690  -0.2539   2.5191  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)    1.37726    0.88776   1.551 0.120807    
gendermale     0.28029    0.23909   1.172 0.241083    
age           -0.04426    0.01825  -2.425 0.015301 *  
yearsmarried   0.09477    0.03221   2.942 0.003262 ** 
childrenyes    0.39767    0.29151   1.364 0.172508    
religiousness -0.32472    0.08975  -3.618 0.000297 ***
education      0.02105    0.05051   0.417 0.676851    
occupation     0.03092    0.07178   0.431 0.666630    
rating        -0.46845    0.09091  -5.153 2.56e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 675.38  on 600  degrees of freedom
Residual deviance: 609.51  on 592  degrees of freedom
AIC: 627.51

Number of Fisher Scoring iterations: 4
```

gender, children, education, occupation may not make significant contribution (you can not reject the hypothesis that the parameters are zero)

can you reject the hypothesis that the parameters are zero?

Infidelity: Build Reduced Model

```
fit.reduced <- glm (ynaffair ~
                      age +
                      yearsmarried +
                      religiousness +
                      rating,
                    data = Affairs,
                    family = binomial(link="logit")
)
summary(fit.reduced)
```

Infidelity: Interpreting Reduced Model

```
Call:
glm(formula = ynaffair ~ age + yearsmarried + religiousness +
    rating, family = binomial(link = "logit"), data = Affairs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6278  -0.7550  -0.5701  -0.2624   2.3998

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.93083    0.61032   3.164 0.001558 **
age           -0.03527    0.01736  -2.032 0.042127 *
yearsmarried   0.10062    0.02921   3.445 0.000571 ***
religiousness -0.32902    0.08945  -3.678 0.000235 ***
rating        -0.46136    0.08884  -5.193 2.06e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 675.38  on 600  degrees of freedom
Residual deviance: 615.36  on 596  degrees of freedom
AIC: 625.36

Number of Fisher Scoring iterations: 4
```

each regression coefficient is statistically significant (p-value < 0.05)

Infidelity: Compare Full & Reduced Models

Because two models are nested (**fit.reduced** is a subset of **fit.full**), use **anova()** function to compare with chi-square version of the test

```
> anova(fit.reduced, fit.full, test = "Chisq")
Analysis of Deviance Table

Model 1: ynaffair ~ age + yearsmarried + religiousness + rating
Model 2: ynaffair ~ gender + age + yearsmarried + children + religiousness +
    education + occupation + rating
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       596     615.36
2       592     609.51  4   5.8474   0.2108
```

*the **non-significant** chi-square value (p-value = 0.21) suggests that the reduced model with fewer predictors fits as well as the full model with nine predictors → use simpler model for further interpretation*

Infidelity: Interpreting model parameters

Model parameters are regression coefficients: $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_q}$

```
> coef(fit.reduced)
(Intercept)           age   yearsmarried religiousness        rating
     1.9308       -0.0353         0.1006       -0.3290       -0.4614
```

In logistic regression, response is modeled as the log(odds) that Y=1.

- The regression coefficients give the change in log(odds) in the response of a unit change in the predictor variable, holding all the other predictors constant.
- log(odds) are hard to interpret; put the results on an odd scale by exponentiation of the coefficient values.

```
> exp(coef(fit.reduced))
(Intercept)           age   yearsmarried religiousness        rating
      6.895         0.965          1.106         0.720         0.630
```

infidelity is increased by a factor of 1.106 for a one-year increase in years married →
a 10-year increase would increase the odds by a factor of $1.106^{10} = 2.7,$ *holding the other predictors constant*

Infidelity: Assessing the impact of predictors on the probability of an outcome

Use predict() function to observe the impact of varying the levels of a predictor variable on the probability of the outcome

```
> test.data <- data.frame(rating=c(1,2,3,4,5),
+                         age = mean(Affairs$age),
+                         yearsmarried = mean(Affairs$yearsmarried),
+                         religiousness = mean(Affairs$religiousness))
```

```
> test.data$prob <- predict(fit.reduced,
+                           newdata = test.data,
+                           type = "response")
> test.data
  rating  age yearsmarried religiousness  prob
1      1 32.5         8.18          3.12 0.530
2      2 32.5         8.18          3.12 0.416
3      3 32.5         8.18          3.12 0.310
4      4 32.5         8.18          3.12 0.220
5      5 32.5         8.18          3.12 0.151
```

Overdispersion: The Expected Variance

The expected variance for data drawn from a binomial distribution is

$$\sigma^2 = np(1-p), \text{ where}$$

n is the number of observations and
$p = P(Y = 1)$, the probability of belonging to the Y=1 group

Overdispersion: when the observed variance of the response variable is larger than what would be expected from a binomial distribution.

- Overdispersion can lead to distorted test standard errors and inaccurate tests of significance
- If overdispersion is present, then use **quasi-binomial** distribution rather than binomial family distribution

Overdispersion: $$\boldsymbol{\phi = \frac{Residual\ deviance}{Residual\ df} \gg 1}$$

$$\boldsymbol{\phi = \frac{615.36}{596} = 1.03}$$

no overdispersion
in infidelity data

Logistic Regression: Extensions

Extension	Issue Addressed	R function / pkg
Robust logistic regression	Data containing outliers and influential observations	glmRob() in pkg=robust
Multinomial logistic regression	Response: multiple unordered categories (married/widowed/divorced)	mlogit() in pkg=mlogit
Ordinal logistic regression	Response: ordered categories (poor/good/excellent credit risk)	lrm() in pkg=rms

Logistic Regression for Classification

Step 1: Estimate the probability that each observation belongs to each class. This is a function of the explanatory variables.

Step 2: Use a cutoff value (often 0.5) on these probabilities to decide which category to put the observation in.

For example, if $P(Y = 1) > 0.5$, then classify the observation as class = 1; if $P(Y = 1) < 0.5$, then classify the observation as class = 0.

You would pick a value besides 0.5 to trade off quantities of interest.

Assess Performance: Classification

Classification/Confusion Matrix

	Predicted Class = 1	Predicted Class = 0
Actual Class = 1	n_{11}	n_{10}
Actual Class = 0	n_{01}	n_{00}

Overall error rate = Estimated misclassification rate = $\frac{n_{10}+n_{01}}{n}$

Overall accuracy = 1 – overall error

Suppose we are interested in class 1.

- *Sensitivity/recall* = ability to detect class of interest = proportion of class 1 classified correctly = $\frac{n_{11}}{n_{10}+n_{11}}$
- *Specificity* = ability to rule out members of "other class" = proportion of class 0 classified correctly = $\frac{n_{00}}{n_{01}+n_{00}}$
- *Precision* = fraction predicted as class of interest that are class of interest = $\frac{n_{11}}{n_{01}+n_{11}}$

Example: Poisson Regression: Seizures

Data Dictionary for the Seizures Data

Trt	Treatment condition
Age	Age in years
Base	# of seizures reported in the baseline 8-week period

What impact does a drug treatment for seizures have on the number of seizures over 8-week period?

Outcome, Y: **Count** (# of siezures)

Example: Poisson Regression: Seizures

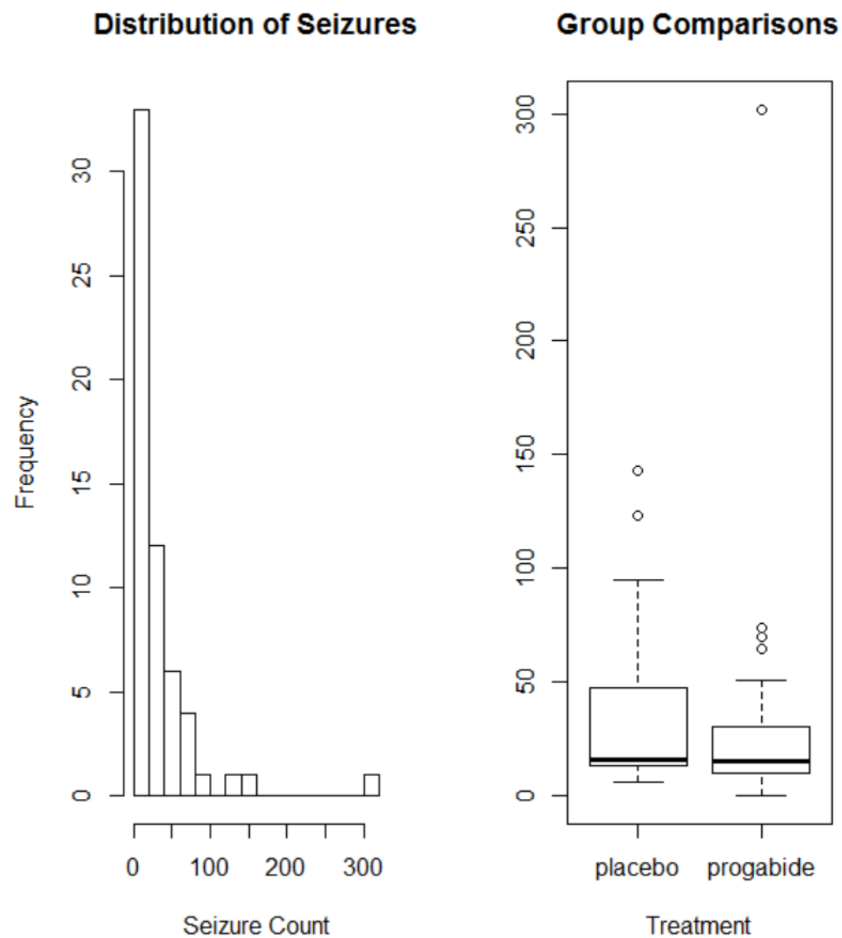
Data Dictionary for the Seizures Data

```
> names(breslow.dat)
 [1] "ID"     "Y1"     "Y2"     "Y3"     "Y4"     "Base"   "Age"
 [8] "Trt"    "Ysum"   "sumY"   "Age10" "Base4"
> dim(breslow.dat)
[1] 59 12
> t(breslow.dat[1,])
      1
ID    "104"
Y1    "5"
Y2    "3"
Y3    "3"
Y4    "3"
Base  "11"
Age   "31"
Trt   "placebo"
Ysum  "14"
sumY  "14"
Age10 "3.1"
Base4 "2.75"
```

What impact does a drug treatment for seizures have on the number of seizures over 8-week period?

Outcome, Y: **Count** (# of siezures)

Seizures: Skewed Predictors & Outliers




```r
hist(sumY, breaks=20, xlab="Seizure Count",
      main="Distribution of Seizures")
boxplot(sumY ~ Trt, xlab="Treatment",
         main = "Group Comparisons")
```

Seizures: Build Poisson Regression Model

```r
fit <- glm(sumY ~ Base + Age + Trt,
                data = breslow.dat,
                family=poisson(link="log"))
summary(fit)
```

```
Call:
glm(formula = sumY ~ Base + Age + Trt, family = poisson(link =
"log"),
    data = breslow.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.0569  -2.0433  -0.9397   0.7929  11.0061

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9488259  0.1356191  14.370  < 2e-16 ***
Base          0.0226517  0.0005093  44.476  < 2e-16 ***
Age           0.0227401  0.0040240   5.651 1.59e-08 ***
Trtprogabide -0.1527009  0.0478051  -3.194   0.0014 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2122.73  on 58  degrees of freedom
Residual deviance:  559.44  on 55  degrees of freedom
AIC: 850.71

Number of Fisher Scoring iterations: 5
```

each regression coefficient is statistically significant (p-value < 0.05)

Seizures: Interpreting Model Parameters

```
> exp(coef(fit))
 (Intercept)          Base           Age Trtprogabide
       7.020         1.023         1.023        0.858
```

one year increase in Age ***multiplies*** *the expected number of seizures by 1.023 → the increased age is associated with higher number of seizures*

Overdispersion: $$\phi = \frac{\textit{Residual deviance}}{\textit{Residual df}} \gg 1$$

$$\phi = \frac{559.54}{55} = 10.17$$

overdispersion in seizure data

Seizures: Overdispersion

Overdispersion: $$\phi = \frac{Residual\ deviance}{Residual\ df} \gg 1$$

$$\phi = \frac{559.54}{55} = 10.17$$

overdispersion in seizure data

Reasons for overdispersion:

- The omission of an important predictor
- State dependence: The probability of a seizure is dependent on other seizures

```
> qcc.overdispersion.test(breslow.dat$sumY,
+                         type = "poisson")

Overdispersion test Obs.Var/Theor.Var Statistic p-value
       poisson data              62.9      3646       0
```

overdispersion in seizure data

Seizures: Dealing with Overdispersion

```r
fit.od <- glm(sumY ~ Base + Age + Trt,
              data = breslow.dat,
              family=quasipoisson())
summary(fit.od)
```

```
Call:
glm(formula = sumY ~ Base + Age + Trt, family = quasipoisson(), 

    data = breslow.dat)

Deviance Residuals: 
   Min       1Q   Median       3Q      Max  
-6.057   -2.043   -0.940    0.793   11.006  

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.94883    0.46509    4.19   0.0001 ***
Base          0.02265    0.00175   12.97   <2e-16 ***
Age           0.02274    0.01380    1.65   0.1051    
Trtprogabide -0.15270    0.16394   -0.93   0.3557    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 11.8)

    Null deviance: 2122.73  on 58  degrees of freedom
Residual deviance:  559.44  on 55  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

taking overdispersion into account leads to insufficient evidence to declare that the drug regimen reduces seizure counts more than receiving a placebo after controlling for baseline seizure rate and age

parameter estimates are identical but standard errors are higher, and p-values for Trt and Age are insignificant