Model	**Equation**	**Distribution**	**Assumptions**
Linear	μi = β0 + β1xi	Yi ~ Normal(μi, σ2)	Linear combination of explanatory terms/ Independence /Normality/ EOV
Poisson Link: Log	log(μi) = β0 + β1xi	Yi ~ Poisson(μi)	Discrete and non-negative distribution / Right-skewed by symmetric when u is big /Variance increases with mean
Logistic Link: Logit	logit(p_i) = b0 + β1xi Var(Yi) = ni pi (1 − pi)	Yi ~ Binomial(ni, pi) Yi ~ Binomial(ni = 1, pi)	Discrete response / Non-constant variance num of successes associated with the ith observation must be an int between 0 and ni
Quasi Poisson	Var(Yi) = ku_i	Yi ~ Quasipoisson(kμ_i)	Multiplying the variance of the corresponding Poisson or logistic regression model by a ‘dispersion parameter’, k.
Negativc Binomial	Var(Yi) = u + u^2 / Ɵ	Yi ~ Negative Binomial(ui,Ɵ)	This distribution assumes that there is a quadratic relationship between the mean and the variance, so Var(Y) > u
Quasi Binomial	Var(Yi) = knipi(1 - pi)	Yi ~ Quasibinomial(ni, pi)	Multiplying the variance of the corresponding Poisson or logistic regression model by a ‘dispersion parameter’, k.
Beta Binomial	Var(Yi) = nipi(1 - pi)(1 + nip-p)	Yi = Beta-binomial(ni, pi, p)	The parameter ρ controls the variance such that ρ = 0 implies no overdispersion and 0 < ρ ≤ 1 accounts for overdispersion

where, for the ith observation, pi is the probability a respondent has a psychiatric illness, GHQi is respondent i's GHQ score, Mi = 1 if the gender is male, otherwise zero. [ensure to include the log off the offset in the equation]

poisson exp(coef): We estimate that for every 1_ increase in_, the expected number of _ is multiplied by _	logistic confint(): a 1 unit increase in _increases the odds _ by a factor between _ to _
poisson exp(100*coef): We estimate that for every 100g increase in the_, the expected number of _ is multiplied by _	logistic exp(coef()): We estimate that, for every 1_increase in _, the odds of _ are multiplied by _.
poisson100*(exp(coef)-1): We estimate that for every 1g increase in_, the expected number of _ increases by _%	logistic exp(10*coef()): We estimate that, for every 10_ increase in _, the odds of having _are multiplied by _.
poisson 100*(exp(100*coef)-1): We estimate that for every 100g increase in the_, the expected number of _ increases by _%	logistic 100*(exp(10*coef()))-1: We estimate that, for every 10_ increase in _, the odds of having _increase by _%.

Calculating Confidence Interval β1 ± 1.96*SE **Z-Stat** = $z = \frac{b}{SE\ of\ b}$

Testing NULL hypothesis T-Statistic = (H1 - H0) / SE **missing Coefficient** Standard Error * Zval

For a T-Stat > 2: Cannot Reject   null hypothesis / In CI = Do not Reject

For a T-Stat < 2 : Reject null hypothesis / Not in CI = Reject

Calculating average value of a explanatory variable (x) with a given probability (80%)

log(p_i / (1 - p_1)) = b0 + b1x + b2$Male_i$ AND p_i = 0.8

$\log(\frac{pi}{1-pi}) = \log((\frac{0.8}{1-0.8}) = \log(4) = b0 + b1x + b2Male_i$ (Rearrange to get x)

$x = \frac{log(4) - b0 - b2Malei}{b1}$

Calculating Deviance & Deviance Residual

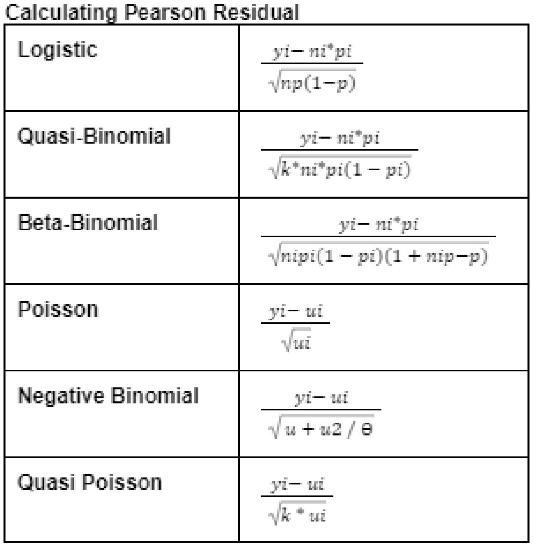
Deviance = 2 (log-likelihood of Saturated - log-likelihood of unsaturated)

$\sqrt{Deviance}$ * d; where d=1 if $\frac{number\ of\ success}{number\ of\ trials}$>p else d=-1

Calculating a Probability of i p = exp(linComb) / (1+ exp(linComb)) OR p_i = (p_i / (1-p_i))

Calculating Raw Residual = y_i - E(Y_i)   ni = total number of observations / pi = (pi / (1-pi))

Calculating Residual Sum Squares (Linear) = $(y_i - u_i)^2$ y_i = number of successes



Calculating Pearson Residual

Logistic	$\frac{yi - ni^*pi}{\sqrt{np(1-p)}}$
Quasi-Binomial	$\frac{yi - ni^*pi}{\sqrt{k^*ni^*pi(1-pi)}}$
Beta-Binomial	$\frac{yi - ni^*pi}{\sqrt{nipi(1-pi)(1+nip-p)}}$
Poisson	$\frac{yi - ui}{\sqrt{ui}}$
Negative Binomial	$\frac{yi - ui}{\sqrt{u + u2 / \theta}}$
Quasi Poisson	$\frac{yi - ui}{\sqrt{k * ui}}$

	POISSON	**LOGISTIC**
DEVIANCE YES	Counts of student attendance in large first-year Stats. Business, Bio, Maths & Psych courses.	How many students PASS in large first-year Stats, Business, Bio, Maths & Psych courses, where each observation is a semester.
DEVIANCE NO	Pass/Fail data for past STATS330 students where each observation is a student.	Data on the count of births per woman (over 16) in Toronto Canada

GROUPED vs UNGROUPED DATA

GROUPED data that has been organised into groups, typically in frequencies		**UN-GROUPED** each individual observation is recorded separately	
Test scores grouped into intervals: B+: 3 students, A: 5 students, A+: 2 students		A list of students' test scores: 85, 90, 78, 92, 88, 76, etc.	
log($odds_i$) = b_0 Y ~ Binomial(n_i, p_i)	\Log(\text{odds_i}) = \beta_0 \\ Y_i \sim \text{Bernoulli}(n_i, p_i)	logit(p_i) = b_0 Y_i ~ Bernoulli(p_i)	\text{logit}(p_i_ = \beta_0 \\ Y_i \sim \text{Bernoulli}(p_i)

Conditions needed to use the Chi-Squared approximation to assess the goodness of fit with the deviance
The distribution of deviance under the null hypothesis is approximately chi-squared if the response of each observation is well approximated by a normal distribution.
This holds for **poisson** random variables with an estimated mean (ui)>=5.
This holds for a **binomial** random variables if the number of trials (n) is large enough

- When pi is close to 0.5 n>=5
- But if pi is close to 1 or 0, ni must be much larger

Offsets
Including the log of the variable by fixing its coefficient to 1. It scales the model to compare response per unit of time.

Log-Likelihood
BIG Log-Likelihood: likelihood of generating the observed data is very low.
SMALL Log-Likelihood: model producing a much higher probability of the observed data.
A log-likelihood closer to zero means that the likelihood (the probability of observing the data given the model) is closer to 1, which represents a near-perfect match between the model and the observed data.

Residual Plots

Observed values minus expected values	Raw Residuals
Observed values minus expected values, divided by Standard Deviation	Pearson residuals
The signed-square root of the deviance contribution for the observation model	Deviance residuals
Residuals with some added jitter for Binomial and Poisson regression (normally distributed when the model is correct, even if there is sparsity)	Randomised Quantile Residuals

Probability	0,1
Logit(p)	-∞,∞
Odds	0,∞
Log(Odds)	-∞,∞
Count	0,∞
Log(Count)	-∞,∞

Diagnosing a GLM

1. Test GoF using deviance statistic (Chi-Square)
2. Inspect Pearson and/or Deviance Residual plots
   a. If there exists patterns (Ice-cream cone) attempt to fix by adding explanatory terms or transformations
   b. If there exists sparsity, fit a Randomised Quantile Residual plot
      i. If RQR appears random scatter, the GLM passes GoF test
      ii. Else, there is a problem with the model
3. If the deviance and residual plots look fine, there is not evidence to suggest the model is inappropriate
4. If the residual plot does not have a pattern but the deviance suggests a lack of fit, the variance of response distribution is probably wrong. (We can try and fit a Quasi-Poisson)

Grouped Binomial	Un-Grouped Binomial	Poisson
Response: missed-booking / all bookings Explanatory: Driving conditions	Response: Wine Quality Explanatory: Acidivy	Response: Number of species Explanatory: Island type

glm(cbind(y, n-y) ~ driving conditions)
logit(pi) = b0 + b1*driving_conditions
Yi ~ Binomial(ni, pi)
	glm(goodQuality ~ acidity)
	logit(pi) = b0 + b1*Acidity
	Yi ~ Binomial(ni=1, pi)
		glm(species ~ Island)
		log(ui) = b0 + b1*Island
		Yi = Poisson(ui)

glm(cbind(y, n-y) ~ 1)
logit(pi) / log(oddsi) / log(pi / 1-pi) = b0
Yi ~ Binomial(1, pi) / Yi ~ Bernoulli(pi)
	glm(goodQuality ~ 1)
	logit(pi) / log(oddsi) / log(pi / 1-pi) = b0
	Yi ~ Binomial(1, pi) / Yi ~ Bernoulli(pi)
		glm(species ~ 1)
		log(ui) = b0
		Yi = Poisson(ui)

Calculating DF
There are 71 observations total.

- enrolment = number of students enrolled at the school
- type = college (C) or university (U)
- nv = the number of violent crimes for that institution for the given year
- enroll1000 = enrolment at the school, in thousands
- region = region of the country (C = Central, MW = Midwest, NE = Northeast, SE = Southeast, and W = West)

Model: fit ← glm(cbind(nv, enrolment-nv) ~ type + region, offset = log(enroll1000), family = 'binomial', data = campus_crime)
Degrees of Freedom = 71 - 6 = 65
Expected counts = enroll1000 * predicted_rate (We model the rate of violent crimes per 1000 students per school)
Yi = Binomial(ni, pi)
log(Oddsi) = b0 + b1TypeUniversityi + b2RegionMWi + b3RegionNEi + b4RegionSEi + b5RegionWi
Where ni = 1 and is the number of successes of the observation i, pi is the probability/proportion of observation i being success, and TypeUniversity=1 if Type=University, and Region = 1 if observation i is from the respective region.

Choosing the best model using Dredge()

- If the difference is less than 2, then both models are similarly supported by the data.
- If the difference is between 2 and 4, the one with the smaller AIC is slightly better supported.
- If the difference is between 4 and 10, the one with the smaller AIC is considerably better supported.
- If the difference is over 10, the one with the larger AIC has essentially no support.