Advanced Statistical Modeling

STA6257

Achraf Cohen - acohen@uwf.edu

Mathematics and Statistics Department
University of West Florida

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful. Essentially, all models are wrong, but some are useful." George Box.

Course

This course covers Statistical Linear Models and aspects related to them.

- We will use *R*. An introduction to *R* will be given and material is posted on Canvas.
- Some packages *GLMsData*, *MASS*, *car*, *AER* will be used.
- Capstone Project.

These slides are being updated if you catch a typo/error please email me at acohen@uwf.edu! Thank you!

Statistical Models

Statistical Models

A statistical model consists of:

- **a random component**: A model of the distribution of the response variable.
- **a systematic component**: The mathematical relationship between Response variable -mean- and Predictors.

$$\mu_i = E[y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \tag{1}$$

OR

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i \tag{2}$$

Regression models

A regression model assumes that the mean response μ_i for Observation i depends on the p explanatory variables (predictors) x_{1i} to x_{pi} via some general function f through a number of regression parameters β_j (for $j = 0, 1, \ldots, p$). Mathematically

$$E[y_i] = \mu_i = f(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_3 x_{pi}) \tag{3}$$

Models in Eq. (3) are *regression models linear in the parameters.* And $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_3 x_{pi}$ is called the *linear predictor.*

Statistical Models

Two types of regression models linear in the parameters are covered in this course:

Linear regression models

The systematic component of a linear regression model assumes the form $E[y_i] = \mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_3 x_{pi}$ and the randomness is assumed to have constant variance σ^2, $y_i \sim N(\mu_i, \sigma^2)$.

Generalized Linear Models (GLM)

The systematic component of a GLM assumes the form $g(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_3 x_{pi}$, $g()$ is called the *link function*. The random component is assumed to follow the Exponential Family of Distributions (EDF).

Linear regression models are a special case of GLM.

Statistical Models

The goal of a statistical model is to accurately represent the important systematic and random components of the data. **But what is the *purpose*** of developing statistical models?
For regression models, there are two major motivations:

- **Prediction**: To produce accurate predictions from new or future data.
- **Understanding** and **interpretation**: To understand how variables relate to each other.

In general, an adequate statistical model balances two criteria:

- **Accuracy**: The model should accurately describe both the systematic and random components.
- **Parsimony**: The model should be as simple as possible

Statistical Models

All models must be used and understood within limitations imposed by **how the data were collected.**

- In an Observational study: we may use elaborate equipment to collect physical measures or may ask subjects to respond to questionnaires, but do not influence the process being observed. In this case, we can only conclude associations between variables.
- In a designed experiment: we may intervene to control the values of the explanatory variables that appear in the data. A well designed randomized experiment allows inferences to be made about cause-and-effect relationships.
- Another feature of data collection that affects conclusions is **the population** from which the subjects are drawn. Conclusions from fitting a statistical model only apply to the population from which the cases are drawn.

Linear regression models

We consider linear regression models for modelling data with a response variable y and p explanatory variables $x_1, x_2, \ldots, x_p$.
A linear regression model consists of the usual two components of a regression model (random and systematic components), with specific forms.

- Random Component: The response y_i have constant variances $var[y_i] = \sigma^2$, for $i = 1, 2, \ldots, n$
- The systematic Component: The expected value of the response is linearly related to the explanatory variables x_j such as:

$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} \tag{4}$$

The regression parameters β_j and σ^2 are unknown and must be estimated from the data. $p + 2$ parameters.

Linearity, Constant variance, and Independence of y_i.

Estimation - Matrix Notation

- **Coefficients** β_j: Ordinary Least Squares (OLS).

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y \tag{5}$$

- **Variance** σ^2:

$$s^2 = \frac{(y - \hat{\mu})^T (y - \hat{\mu})}{n - p - 1} = \frac{RSS}{n - p - 1} \tag{6}$$

- **Variance of** $\hat{\beta}$: The covariance matrix is

$$var[\hat{\beta}] = \sigma^2 (X^T X)^{-1} \tag{7}$$

  An estimate is $\widehat{var}[\hat{\beta}] = s^2 (X^T X)^{-1}$

- **Variance of Fitted values**: Given data x_{new}, the best estimate of the mean response is $\hat{\mu}_{new} = x_{new} \hat{\beta}$. The variance of $\hat{\mu}_{new}$ is:

$$var[\hat{\mu}_{new}] = \sigma^2 x_{new} (X^T X)^{-1} x_{new}^T \tag{8}$$

Normal Linear regression models

Normality is only important to develop tests and CI that are valid for small sample size.

Inference - Normal

- **Distribution of** β_j:

$$\hat{\beta}_j \sim N(\beta_j, var[\hat{\beta}_j]) \tag{9}$$

- **Hypothesis tests for** β_j: $H_0 : \beta_j = \beta_j^*$ against a one-sided alternative ($H_A : \beta_j > \beta_j^* or \beta_j < \beta_j^*$) or a two-sided alternative ($H_A : \beta_j \neq \beta_j^*$). The hypothesized value β_j^* (usually zero). The statistic

$$T = \frac{\hat{\beta}_j - \beta_j^*}{se(\hat{\beta}_j)} \sim t_{n-p-1} \tag{10}$$

- **CI for** β_j: $100(1-\alpha)\%$ confidence intervals for each estimate:

$$\hat{\beta}_j \pm t_{(\alpha/2, n-p-1)} se(\hat{\beta}_j), \tag{11}$$

- **CI for** μ: The $100(1-\alpha)\%$ confidence intervals for the fitted value:

$$\hat{\mu}_{new} \pm t_{(\alpha/2, n-p-1)} se(\hat{\mu}_{new}) \tag{12}$$

ANOVA for regression models

It is of interest to know whether the explanatory variables are useful predictors of the responses. This question can be answered statistically by testing whether the regression sum of squares *SSReg* is larger than *RSS* residuals sum of squares. The ANOVA identity can answer this question.

- **F statistic**:

$$F = \frac{SSReg/(p)}{RSS/(n-p-1)} = \frac{MSReg}{MS_E} \sim F_{(p,n-p-1)} \tag{13}$$

  Note that MS_E is equal to s^2 (Eq.(6)), the unbiased estimator of σ^2.

- **The Coefficient of Determination** R^2: The proportion of the total variation explained by the regression:

$$R^2 = 1 - \frac{RSS}{SS_T} \tag{14}$$

- $\boldsymbol{R^2_{adjusted}}$: The proportion of the total variation explained by the regression adjusted for the number of explanatory variables:

$$R^2_{adjusted} = 1 - \frac{RSS/(n-p-1)}{SS_T/(n-1)} \tag{15}$$

ANOVA to compare 2 nested models

Consider:

Model A: $\mu_A = \beta_0 + \beta_1 x_1 + \beta_4 x_4$

Model B: $\mu_B = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

Model A is nested in Model B.

- Comparing these models, we want to know whether the more complex Model B is necessary, or the simpler Model A will suffice.
- Formally, $H_0 : \beta_3 = \beta_4 = 0$ (2 models are equivalent) and H_A : both $\beta_3 \neq 0$ and $\beta_4 \neq 0$

$$F = \frac{(RSS_A - RSS_B)/(p_b - p_A)}{RSS_B/dfres_B} \sim F_{(p_b - p_A, dfres_B = n - p_b - 1)} \tag{16}$$

> anova(Model A, Model B)

Non-nested regression models

First, recall that the two criteria for selecting a statistical model are accuracy and parsimony.

AIC - Akaike's Information Criterion

$$AIC = n \log(RSS/n) + 2(p+1) \tag{17}$$

smaller values of the AIC (closer to $-\infty$) represent better models. The term $2(p+1)$ is called the *penalty*. **AIC focuses more on models with good predictions.**

BIC - Bayesian Information Criterion

$$BIC = n \log(RSS/n) + log(n)(p+1) \tag{18}$$

The BIC is inclined to select lower dimensional (more parsimonious) models than is AIC. Smaller values of BIC (closer to $-\infty$) represent better models. The term $\log(n)(p+1)$ is called the *penalty*. **BIC is a trade-off between prediction and Interpretation.**

Model Selection: *Forward*, *Backward*, and *stepwise* regressions.

The residuals are the the information that is left not explained by the model. They are useful to investigate the model's lack of fit.

Raw Residuals

$$r_i = y_i - \hat{\mu}_i \tag{19}$$

Modified residuals

$$r_i^* = \frac{y_i - \hat{\mu}_i}{\sqrt{1 - h_i}} \tag{20}$$

where h_i is the *leverage* (hat-values $h_i = h_{ii} = \sum_{j=1}^{n} h_{ij}$, Idempotent matrix) which y_i has in estimating its own fitted value. An interesting interpretation of $(r_i^*)^2$ is the reduction in the RSS that results when Obs. i is omitted.

Standardized residuals

$$r_i' = \frac{r_i^*}{s} = \frac{y_i - \hat{\mu}_i}{s\sqrt{1 - h_i}} \sim t_{n-p-1} \tag{21}$$

Why all these types of residuals?

The residuals are the the information that is left not explained by the model. They are useful to investigate the model's lack of fit.

Raw Residuals

$$r_i = y_i - \hat{\mu}_i \tag{19}$$

Modified residuals

$$r_i^* = \frac{y_i - \hat{\mu}_i}{\sqrt{1 - h_i}} \tag{20}$$

where h_i is the *leverage* (hat-values $h_i = h_{ii} = \sum_{j=1}^{n} h_{ij}$, Idempotent matrix) which y_i has in estimating its own fitted value. An interesting interpretation of $(r_i^*)^2$ is the reduction in the RSS that results when Obs. i is omitted.

Standardized residuals

$$r_i' = \frac{r_i^*}{s} = \frac{y_i - \hat{\mu}_i}{s\sqrt{1 - h_i}} \sim t_{n-p-1} \tag{21}$$

Why all these types of residuals? $var[r_i] = \sigma^2(1 - h_i), var[r_i^*] = \sigma^2$

Model Diagnostics

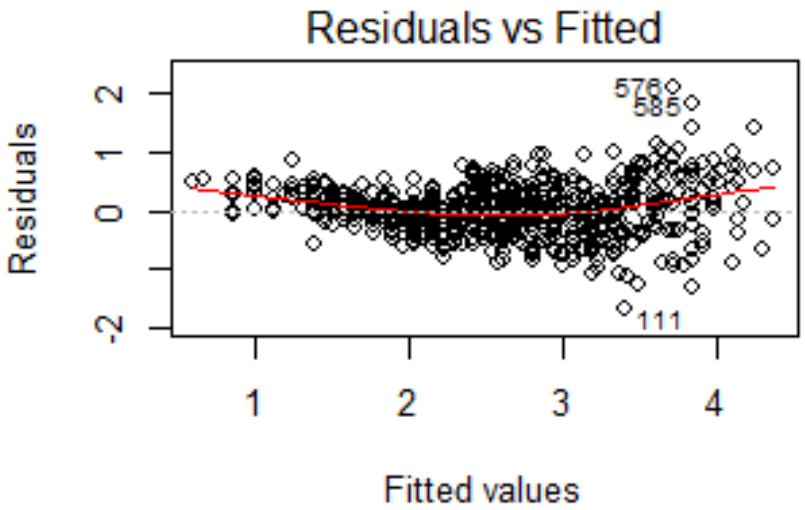
A plot of **Standardized** residuals against a covariate x_j can more easily detect deviations from linearity, because the linear effects of all the explanatory variables have been removed.

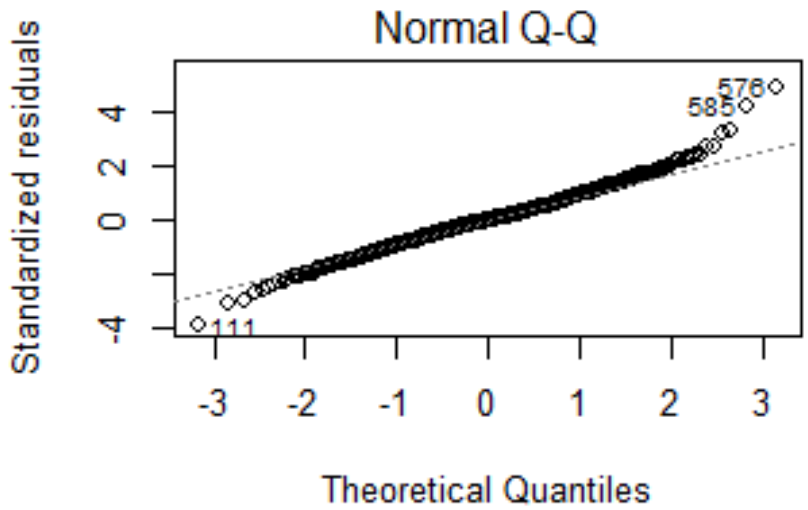
Plot Residuals vs. x_j Linearity

Plot Residuals vs. Fitted values Constant Variance

Quantile-Quantile Normality

Lag plot OR AutoCorrelation Function Dependence over Time





Model Diagnostics

In this section we discuss methods to identify problems with individual observations - **Outliers and Influential Observations** -.

Outliers are observations inconsistent with the rest of the data set. As a guideline, potential outliers might be flagged as observations with standardized residual $| r' | > 2.5$.

Influential observations are observations that substantially change the fitted model when omitted from the data set.

Cook's distance

$$D_i = \frac{(r_i')^2 h_i}{(p+1)(1-h_i)} \approx F_{p+1,n-p-1} \tag{22}$$

A useful rule-of-thumb is that observations with $D > 1$ may be flagged as potentially influential.

DFFITS

$$DFFITS_i = \frac{\hat{\mu}_i - \hat{\mu}_{i(i)}}{s_{(i)}} = r'' \sqrt{\frac{h_i}{1-h_i}} \tag{23}$$

measures how much the fitted value of Observation i changes between the model fitted with all the data and the model fitted when Observation i is omitted.

Influential observations are observations that substantially change the fitted model when omitted from the data set.

DFBETAS

$$DFBETAS_i = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{se(\hat{\beta}_{j(i)})} \tag{24}$$

measures how much the estimates of each individual regression coefficient change between the model fitted using all observations and the model fitted with Observation i omitted.

Covariance ratio - CR

$$CR = \frac{1}{1-h}\left(\frac{n-p}{n-p-1+(r'')^2}\right)^p \tag{25}$$

measures the increase in uncertainty about the regression coefficients when Observation i is omitted.
Where r'' is the *Studentized* residual. $r_i'' = \frac{y_i - \hat{\mu}_{i(i)}}{s_{(i)}\sqrt{1-h_i}} \sim t_{n-p-1}$

Model Diagnostics

The **R** function **influence.measures()** produces a table of the influence measures *dfbetas*, *dffits*, *cr* and *D*, plus the leverages *h*. Observations identified as influential with respect to any of these statistics (or having high leverage in the case of h) are flagged with a * according to the following criteria:

- Leverages h: Observations are declared high leverage if h > $3(p+1)/n$.
- Cook's distance *D*: Observation *i* is declared influential when *D* exceeds the 50*th* percentile of the $F_{p+1,n-p-1}$.
- DFBETAS: Observation *i* is declared influential when $\mid DFBETAS_i \mid > 1$.
- DFFITS: Observation *i* is declared influential when $\mid DFFITS_i \mid > \frac{3}{\sqrt{(p+1)/(np-1)}}$
- Covariance ratio CR: Observation *i* is declared influential when $CR_i > \frac{3(p+1)}{(np-1)}$.

Model building

We shall identify some specific problems and some remedies to ameliorate the fitted model.

- **Variance non-constant**: Stabilizing by transforming the response using Box-Cox transformation.
- **Nonlinear** relationship between y and x can be fixed by a transformation of x.
- **Dependence** between responses such as repeated measures, time series, or longitudinal studies. Mixed models and Generalized LS should be used in this case.
- When the cause of an **outlier** cannot be identified, a strategy to evaluate the influence of the outlier is to fit the model to the data with and without the outlier.
- **Collinearity**: Occurs when some of the covariates are highly correlated with each other. It can be accommodated by i) Omit some explanatory variables from the analysis, ii) Combine explanatory variables in the model, iii) use ridge regression.

Maximum Likelihood Estimation

- Least-squares is an appropriate criterion for fitting regression models to response data that are *approximately normally distributed*.
- A more general estimation methodology is needed when data fail to be approximately normally distributed, such as:
    1. Proportions of a total number of counts - **Binomial** -
    2. Discrete count data may be modelled by the **Poisson** or **negative binomial** distributions.
    3. Positive Continuous data may be modelled by the **gamma** and **inverse Gaussian** distributions
    4. Positive data with exact zeros may be modelled by a special case of the **Tweedie** distributions

The idea of maximum likelihood is to choose those estimates for the unknown parameters that maximize the probability density of the observed data.

Maximum Likelihood Estimation

Suppose X is a random variable with probability distribution $f(x, \theta)$, where θ is an unknown parameter. Let $x_1, x_2, \ldots, x_n$ be a random sample of n observations. The joint probability distribution of the sample is given by $\prod_{i=1}^{n} f(x_i, \theta)$. We can write the likelihood function as:

$$L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta) \tag{26}$$

The maximum likelihood estimator (MLE) of θ is the value of θ that maximizes the likelihood function $L(x_1, x_2, \ldots, x_n; \theta)$. We usually maximize the log-likelihood:

$$l(x_i, \theta) = \log L(x_1, x_2, \ldots, x_n; \theta) = \sum_{i=1}^{n} \log f(x_i, \theta) \tag{27}$$

Maximum Likelihood Estimation

Score Equations: The derivative of the log-likelihood is the score function

$$\mathcal{U}(\theta) = \frac{dl}{d\theta} \tag{28}$$

and the score equations is

$$\mathcal{U}(\theta) = 0 \tag{29}$$

Information The second derivatives of the log-likelihood quantify the amount of information available for estimating parameters. Intuitively, It is it large then $\mathcal{U}(\theta)$ is changing rapidly near the MLE and a peak of the log-likelihood is very sharp, which implies that the estimate is well defined.

$$\mathcal{I}(\theta) = E\Big[-\frac{d^2 l}{d\theta^2}\Big] = E\Big[-\frac{d\mathcal{U}}{d\theta}\Big] \tag{30}$$

where $-\frac{d\mathcal{U}}{d\theta}$ is called the **observed information**. And $\mathcal{I}(\theta)$ is called **Fisher information**.

Maximum Likelihood Estimation

Maximum likelihood estimators have many appealing properties, which we state in this section without proof. Consider $\hat{\theta}$ is the MLE of θ.

- MLEs are **invariant**. This means that if $f(\theta)$ is a one-to-one function of θ, then $f(\hat{\theta})$ is the MLE of $f(\theta)$.
- MLEs are **asymptotically unbiased**. This means that $E[\hat{\theta}] = \theta$ as $n \longrightarrow \infty$. For small samples, the bias may be substantial. In some situations (such as the parameter estimates β_j in normal linear regression models), the MLE is unbiased for all n.
- MLEs are **asymptotically efficient**. This means that no other asymptotically unbiased estimator exists with **a smaller variance**. Furthermore, if an efficient estimator of θ exists, then it must be asymptotically equivalent to $\hat{\theta}$.
- MLEs are **consistent**. This means that the MLE converges to the true value of θ for increasing n: $\hat{\theta} \xrightarrow{p} \theta$.
- MLEs are **asymptotically normally distributed**.

$$\hat{\theta} \sim N(\theta, 1/\mathcal{I}(\theta)) \tag{31}$$

as $n \longrightarrow \infty$.

Hypothesis Testing: Large Sample

We are interested in the hypotheses that:

$$H_0 : \quad \theta = \theta_0$$
$$H_1 : \quad \theta \neq \theta_0$$

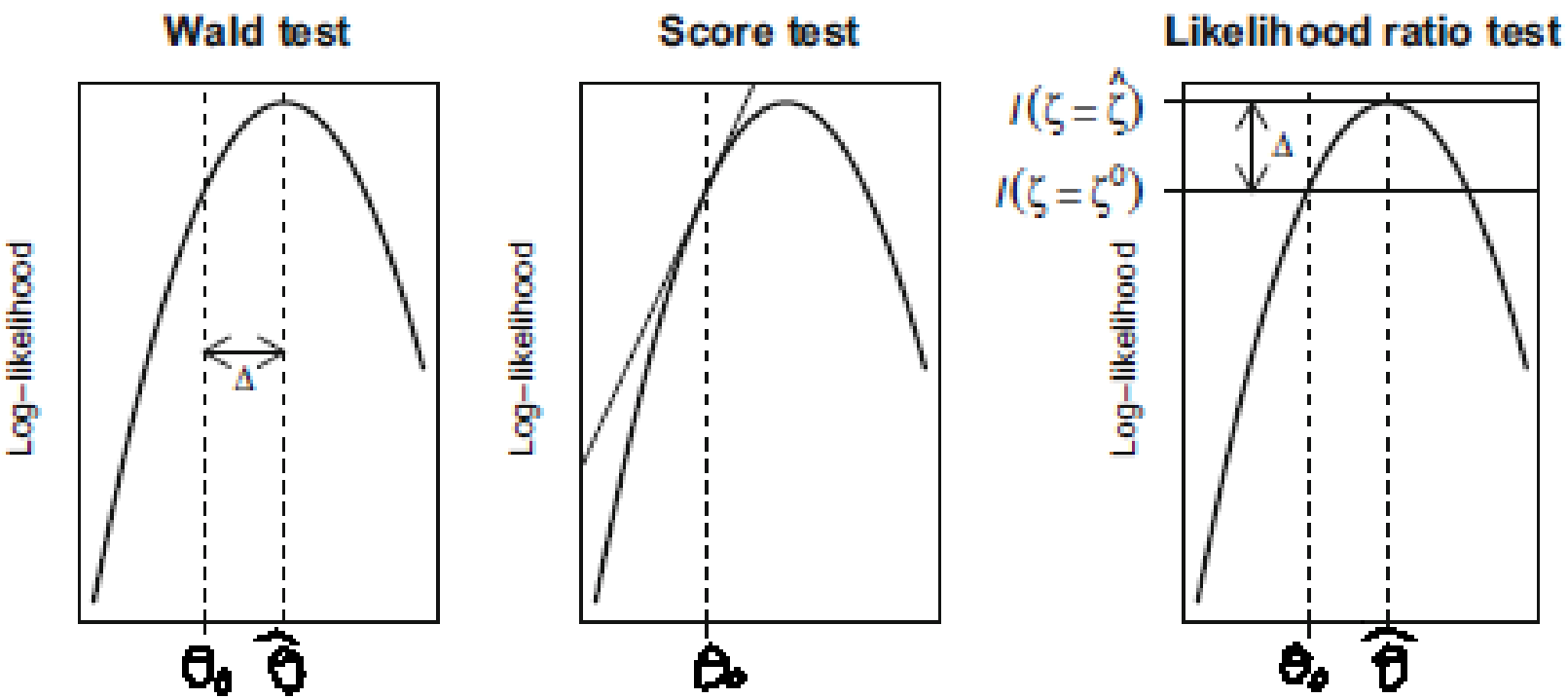


Figure: Adapted from [2]

Hypothesis Testing: Large Sample $n \longrightarrow \infty$.

Wald Test Examine the distance.

$$W = \frac{(\hat{\theta} - \theta_0)^2}{var[\hat{\theta}]} \overset{H_0 isTRUE}{\sim} \chi^2_1 \tag{32}$$

We often use the square root of the W, which follows $N(0, 1)$

Score test Examine the slope of the log-likelihood near $\hat{\theta}$

$$S = \frac{\mathcal{U}(\theta_0)^2}{\mathcal{I}(\theta_0)} \overset{H_0 isTRUE}{\sim} \chi^2_1 \tag{33}$$

The slope of log-likelihood is 0 at $\hat{\theta}$ ($\mathcal{U}(\hat{\theta}) = 0$), so if the log-likelihood at θ_0 is near 0, then θ_0 is near $\hat{\theta}$. Note that $var[\mathcal{U}(\theta)] = \mathcal{I}(\theta)$.

Likelihood Ratio Test is based on the distance between the maximum possible value of the log-likelihood (evaluated at $\hat{\theta}$) and the likelihood evaluated at θ_0

$$L = 2(l(\hat{\theta}) - l(\theta_0)) \overset{H_0 isTRUE}{\sim} \chi^2_1 \tag{34}$$

Generalized Linear Models: Structure

Three components of the GLMs.

Random Component What probability distribution is appropriate? OR knowledge of how the variance changes with the mean?

Systematic Component Linear Predictor

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{35}$$

Link function Component $g(.)$ links the mean μ to the *linear predictor*

$$\boldsymbol{g}(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{36}$$

Generalized Linear Models: Random Component

GLMs assumes the response come from a distribution that belongs to a family of distributions called ***exponential dispersion model*** or **EDMs**. It includes *Normal*, *Gamma*, *Binomial*, *Poisson*, *Negative Binomial*, and *Inverse Gaussian* distributions.

Definition of EDMs

The EDM family has a probability function of the form:

$$\mathcal{P}(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{y\theta - k(\theta)}{\phi}\right\}, \tag{37}$$

- θ is called the *canonical* parameter.
- $k(\theta)$ is the *cumulant* function.
- $\phi > 0$ is the *dispersion* parameter.
- $a(y,\phi)$ is a normalizing function that ensures that (37) is a probability function.
- The mean μ is a known function of the canonical parameter θ.

Examples:

$$\mathcal{P}(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{y\theta - k(\theta)}{\phi}\right\}, \tag{38}$$

1. Normal density probability function

$$\mathcal{P}(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{y^2}{2\sigma^2}\right\}\exp\left\{\frac{y\mu - (\mu^2/2)}{\sigma^2}\right\} \tag{39}$$

2. Poisson mass probability function:

$$\mathcal{P}(y;\mu) = \frac{1}{y!}\exp\left\{y\log\mu - \mu\right\} \tag{40}$$

Generating Functions

The *Moment Generating Function*, denoted *mgf(t)*, for some random variable Y with probability function $\mathcal{P}(y)$ is:

$$mgf(t) = E[e^{tY}] = \begin{cases} \int \mathcal{P}(y)e^{ty}dy & \text{for Y continuous} \\ \sum \mathcal{P}(y)e^{ty} & \text{for Y discrete} \end{cases} \tag{41}$$

for all values of t for which the expectation exists. The *cumulant generating function* (or cgf) is then defined as

$$K(t) = \log mgf(t) = \log E[e^{ty}], \tag{42}$$

The cgf is used to derive the cumulants of a distribution, such as the **mean (first cumulant, K_1)** and the **variance (second cumulant, K_2)**. The *rth* cumulant, K_r, is

$$k_r = \left. \frac{d^r K(t)}{dt^r} \right|_{t=0} \tag{43}$$

The Moment Generating Function of EDMs:

$$mgf(t) = \exp\left\{\frac{k(\theta + t\phi) - k(\theta)}{\phi}\right\} \tag{44}$$

The Cumulant Function of EDMs:

$$K(t) = \frac{k(\theta + t\phi) - k(\theta)}{\phi} \tag{45}$$

Then the *rth* cumulant of an EDM is:

$$k_r = \phi^{r-1}\frac{d^r k(\theta)}{d\theta^r} \tag{46}$$

Generalized Linear Models: Random Component

Consider $Y \sim EDM(\mu, \phi)$

Mean

$$E[Y] = \mu = k'(\theta) \qquad \text{(r=1 in (46))} \tag{47}$$

Variance

$$Var[Y] = \phi k''(\theta) \qquad \text{(r=2 in (46))} \tag{48}$$

We should observe that:

$$k''(\theta) = \frac{d^2 k(\theta)}{d\theta^2} = \frac{d\mu}{d\theta} > 0$$

That means μ must be monotonically increasing function of θ. Define:

$$V(\mu) = \frac{d\mu}{d\theta} \tag{49}$$

This is **called the variance function.** Then $Var[Y] = \phi V(\mu)$

The Variance Function

The variance function $V(\mu)$ *uniquely* determines the distribution within the class of EDMs (function of the cumulant function).

$$Var[Y] = \phi V(\mu) \tag{50}$$

- This is a very important result because if the mean-variance relationship can be established for a given data set then the corresponding EDM is uniquely determined.
- The equation (50) states that the variance of an EDM depends on the mean, except for the normal distribution where the variance does not depend on the mean ($V(\mu) = 1$).

Let's define the following function of μ:

$$t(y, \mu) = y\theta - k(\theta) \tag{51}$$

And

$$\frac{\partial t(y, \mu)}{\partial \theta} = y - \frac{dk(\theta)}{d\theta} = y - \mu \tag{52}$$

$$\frac{\partial^2 t(y, \mu)}{\partial \theta^2} = -k''(\theta) = -V(\mu) < 0 \tag{53}$$

The second derivative is negative and the first derivative is zero at $y = \mu$. Then $t(y, \mu)$ must have a unique maximum with respect to μ at $\mu = y$.

The Unit Deviance

The unit deviance is defined as follows:

$$d(y, \mu) = 2\{t(y, y) - t(y, \mu)\} \tag{54}$$

- $d(y, \mu) = 0$ at $y = \mu$
- $d(y, \mu) > 0$
- It increases as μ moves away from y in either direction.
- $d(y, \mu)$ can be seen as a **distance measure** between y and μ.

Generalized Linear Models: The Unit Deviance

The unit deviance is defined as follows:

$$d(y, \mu) = 2\{t(y, y) - t(y, \mu)\} \approx \phi\chi_1^2 \tag{55}$$

The unit deviance is always chi-square for the *Normal* and *inverse Gaussian* distributions. For other common EDMs the unit deviance is roughly chi-square with the correct expected value when:

- Binomial distribution: $m\mu \geq 3$ and $m(1-\mu) \geq 3$
- Poisson distribution: $\mu \geq 3$
- Gamma distribution: $\phi \leq 1$

We rewrite the EDM as follows:

$$\mathcal{P}(y; \mu, \phi) = A(y, \phi) \exp\left\{\frac{-d(y, \mu)}{2\phi}\right\}, \tag{56}$$

Generalized Linear Models: The Unit Deviance

This table shows common EDMs:

EDM	$V(\mu)$	$\kappa(\theta)$	θ	ϕ	$d(y,\mu)$	S	Ω	Θ
Normal	1	$\theta^2/2$	μ	σ^2	$(y-\mu)^2$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$
Binomial	$\mu(1-\mu)$	$\frac{\exp\theta}{1+\exp\theta}$	$\log\frac{\mu}{1-\mu}$	$\frac{1}{m}$	$2\left\{y\log\frac{y}{\mu}+(1-y)\log\frac{1-y}{1-\mu}\right\}$	$\frac{0,1,\dots m}{m}$	$(0,1)$	$\mathbb{R}$
Negative binomial	$\mu+\frac{\mu^2}{k}$	$-\log(1-\exp\theta)$	$\log\frac{\mu}{\mu+k}$	1	$2\left\{y\log\frac{y}{\mu}-(y+k)\log\frac{y+k}{\mu+k}\right\}$	$\mathbb{N}_0$	$\mathbb{R}^+$	$\mathbb{R}^-$
Poisson	μ	$\exp\theta$	$\log\mu$	1	$2\left\{y\log\frac{y}{\mu}-(y-\mu)\right\}$	$\mathbb{N}_0$	$\mathbb{R}^+$	$\mathbb{R}$
Gamma	μ^2	$-\log(-\theta)$	$-\frac{1}{\mu}$	ϕ	$2\left\{-\log\frac{y}{\mu}+\frac{y-\mu}{\mu}\right\}$	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}$
Inverse Gaussian	μ^3	$-\sqrt{-2\theta}$	$-\frac{1}{2\mu^2}$	ϕ	$\frac{(y-\mu)^2}{\mu^2 y}$	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}_0^-$
Tweedie ($\xi\le 0$ or $\xi\ge 1$)	μ^ξ	$\frac{\{(1-\xi)\theta\}^{(2-\xi)/(1-\xi)}}{2-\xi}$ for $\xi\neq 2$	$\frac{\mu^{1-\xi}}{1-\xi}$ for $\xi\neq 1$	ϕ	$2\left\{\frac{\max(y,0)^{2-\xi}}{(1-\xi)(2-\xi)}-\frac{y\mu^{1-\xi}}{1-\xi}+\frac{\mu^{2-\xi}}{2-\xi}\right\}$ for $\xi\neq 1,2$	$\xi<0$: $\mathbb{R}$ $1<\xi<2$: $\mathbb{R}_0^+$ $\xi>2$: $\mathbb{R}^+$	$\mathbb{R}^+$ $\mathbb{R}^+$ $\mathbb{R}^+$	$\mathbb{R}_0^+$ $\mathbb{R}^-$ $\mathbb{R}_0^-$

Generalized Linear Models: The Total Deviance

An overall measure of the distance between all the y_i and all the μ_i can be defined as

$$D(y, \mu) = \sum_{i=1}^{n} d(y_i, \mu_i) \sim \phi \chi^2_n \tag{57}$$

This is called the **deviance function** and its value called **the deviance.** We rewrite the EDM as follows:

$$\mathcal{P}(y; \mu, \phi) = A(y, \phi) \exp\left\{ \frac{-d(y, \mu)}{2\phi} \right\}, \tag{58}$$

Therefore log-Likelihood Function:

$$l(y; \mu, \phi) = \sum_{i=1}^{n} \log A(y_i, \phi) - \frac{D(y, \mu)}{2\phi} \tag{59}$$

Generalized Linear Models: The Systematic Component - Link function

GLMs assume a systematic component where the *linear predictor:*

$$\boldsymbol{g}(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{60}$$

- The link function $g(.)$ is a *monotonic*, *differentiable* function relating μ_i to the linear predictor η_i.
- Monotonicity ensures that any value of η_i is mapped to only one possible value of μ_i.
- Differentiability is required for estimation.
- The ***canonical* link function** is a special link function, the function $g(\mu)$ such that $\eta = \theta = g(\mu)$.

Examples: Normal and Poisson.

Generalized Linear Models: The Systematic Component - Offsets

The *linear predictor:*

$$\boldsymbol{g}(\mu_i) = O_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{61}$$

where O_i are the *offsets* terms. They can be seen as terms where β_j are known a priori.

- For example the number of births μ_i (counts data) in City i depends on the population pop_i. Poisson model:

$$\boldsymbol{\log}(\mu_i/pop_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\boldsymbol{\log}(\mu_i) = \boldsymbol{\log}(pop_i) + \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Generalized Linear Models: The Link function

This table shows common link functions:

Link function	gaussian	binomial and quasibinomial	poisson and quasipoisson	Gamma	inverse.gaussian	quasi
identity	★		✓	✓	✓	★
log	✓		★	✓	✓	✓
inverse	✓			★	✓	✓
sqrt			✓			✓
1/mu^2					★	✓
logit		★				✓
probit		✓				✓
cauchit		✓				
cloglog		✓				✓
power			star = canonical			✓

Generalized Linear Models: Estimation

The GLMs assume a specific probability distribution for the responses from the EDM family, then MLE methods are used for parameter estimation.

$$\mathcal{P}(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{y\theta - k(\theta)}{\phi}\right\}, \tag{62}$$

Consider we observe $y_1, y_2, y_3, ..., y_n$ and $y_i \sim \mathcal{P}(y_i;\theta_i,\phi)$, then:

$$l(y_i;\theta_i,\phi) = \sum_{i=1}^{n} \log \mathcal{P}(y;\theta_i,\phi)$$

$$l(y_i;\theta_i,\phi) = \sum_{i=1}^{n} \log a(y_i,\phi) + \sum_{i=1}^{n} \frac{y_i\theta_i - k(\theta_i)}{\phi}$$

The GLMs:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{63}$$

Then the score equations:

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0, \qquad j = 0, 1, \ldots, p \tag{64}$$

Computing estimates of β_j

The *Fisher scoring algorithm* provides an effective method for computing MLEs $\hat{\beta}_j$. It is an iterative algorithm. Does not depend on knowing ϕ.

- MLE can be shown to be equivalent to minimizing the Total deviance.

Generalized Linear Models: Estimation of ϕ

Although knowledge of ϕ was not required for estimating the β_j, it will be required for hypothesis testing and confidence intervals

- The most common models for which ϕ is known are binomial ($\phi = 1/n$; n is #trials) and Poisson ($\phi = 1$) EDMs.
- *MLE*: $\hat{\phi}_{MLE}$ is biased unless $n >> p$
- The *modified profile log-likelihood* (MPL): The MPL estimator is a consistent estimator and is approximately unbiased, even in quite small samples. Inconvenient to compute.
- *Mean Deviance* Estimator: $\hat{\phi} = \frac{D(y,\hat{\mu})}{n-p-1}$. It is convenient to compute.
- *Pearson* Estimator: $\hat{\phi} = \sum_{i=1}^{n} \frac{(y_i-\hat{\mu}_i)/V(\hat{\mu}_i)}{n-p-1}$. Convenient and always appro. unbiased. Few assumptions. Pearson estimator tends to be more variable (less precise) but less biased than the mean deviance estimator.

Which one to use?

Generalized Linear Models: Estimation of ϕ

Although knowledge of ϕ was not required for estimating the β_j, it will be required for hypothesis testing and confidence intervals

- The most common models for which ϕ is known are binomial ($\phi = 1/n$; n is #trials) and Poisson ($\phi = 1$) EDMs.
- *MLE*: $\hat{\phi}_{MLE}$ is biased unless $n >> p$
- The *modified profile log-likelihood* (MPL): The MPL estimator is a consistent estimator and is approximately unbiased, even in quite small samples. Inconvenient to compute.
- *Mean Deviance* Estimator: $\hat{\phi} = \frac{D(y,\hat{\mu})}{n-p-1}$. It is convenient to compute.
- *Pearson* Estimator: $\hat{\phi} = \sum_{i=1}^{n} \frac{(y_i-\hat{\mu}_i)/V(\hat{\mu}_i)}{n-p-1}$. Convenient and always appro. unbiased. Few assumptions. Pearson estimator tends to be more variable (less precise) but less biased than the mean deviance estimator.

Which one to use? *glm()* function in *R* uses the Pearson estimator.

ϕ is known.
Same hypothesis tests are applied to GLM as we have seen before (see Slide 25). 3 tests are available: *Wald*, ***Score***, and Likelihood Ratio (LR).

- Wald statistic $\sim N(0, 1)$
- To compare nested models A and B use LR test:

$$L = 2(l_B - l_A) = \frac{D(y, \hat{\mu}_A) - D(y, \hat{\mu}_B)}{\phi} \overset{H_0 \text{is} TRUE}{\sim} \chi^2_{p_B - p_A} \tag{65}$$

- *Analysis of Deviance* tables: generalization of ANOVA tables. *anova()* function in R. The argument *test="Chisq"* must be specified as well as ϕ.
- **Score test** can calculated using *glm.scoretest()* in package **statmod**. Statistic $\sim N(0, 1)$

ϕ is unknown.

- Wald statistic $\sim t_{n-p-1}$
- To compare nested models A and B use LR test:

$$F = \frac{D(y, \hat{\mu}_A) - D(y, \hat{\mu}_B)}{(p_B - p_A)s^2} \overset{H_0 \text{ is TRUE}}{\sim} F_{p_B - p_A, n - p_B} \tag{66}$$

  In R, *lr.test{mdscore}*.

- *Analysis of Deviance* tables: *anova()* function in R. The argument *test="F"* must be specified as well as ϕ.
- **Score test** can calculated using *glm.scoretest()* in package **statmod**. Statistic $\sim t_{n-p-1}$.

Generalized Linear Models: Goodness-of-Fit

Goodness-of-Fit (GoF) tests determine whether the current linear predictor already includes enough explanatory variables to fully describe the systematic trends in the data. This sort of test is only possible when ϕ is known because it requires a known distribution of residuals.

- A GoF compares your model with a *saturated* model (has many explanatory variables as data points).
- The fitted values of the saturated model are all equal to the observed data points $\hat{\mu}_i = y_i$,
- If **the test is rejected** then there is evidence that the model not adequate, which means that the systematic component does not explain everything that can be explained.

Deviance GoF test The residual deviance for the saturated model is zero $D(y, \hat{\mu}_{saturated})$. Therefore, the likelihood-ratio for the current vs. saturated is the residual deviance (Eq. 59):

$$logLike_s - logLike_o \approx Deviance_o - Deviance_s \approx \chi^2_{n-p-1} \quad (67)$$

Pearson GoF test Pearson statistic

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \approx \chi^2_{n-p-1} \quad (68)$$

Generalized Linear Models: Diagnostics

Diagnostics are done based on residuals

Response Residuals are the distances $y_i - \hat{\mu}_i$ and they are not adequate for GLMs.

Pearson Residuals to handle non-constant variance in EDMs:

$$r_p = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}} \tag{69}$$

In R *resid(model,type="pearson")*.

Deviance Residuals are defined as:

$$r_D = sign(y - \hat{\mu})\sqrt{d(y, \hat{\mu})} \tag{70}$$

($sign(x) = 1$ if $x > 0$; -1 if $x < 0$; and 0 if $x = 0$). In R *resid(model)*.

Quantile Residuals are an alternative, which are exactly normally distributed. **qresid(model){statmod}**. Especially for discrete EDMs.

The Pearson and deviance residuals have **approximate normal distributions**, with ***the deviance residuals more likely to be more normally distributed*** *than the Pearson residuals.*

Models for **Proportions**

The outcome of many studies is a proportion y of a total number m:

- the proportion of individuals having a disease
- the proportion of voters who vote in favour of a particular election candidate

The binomial distribution is an EDM where:

- $V(\mu) = \mu(1-\mu)$
- $\phi = 1$, given m is known
- $\theta = \log \frac{\mu}{1-\mu}$ is the link canonical function.

The GLMs:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{71}$$

where $g(.)$ can be:

Logistic $\log \frac{\mu}{1-\mu}$. log-Odds.

Probit $\Phi^{-1}(.)$; Φ is standard normal CDF.

Complementary log-log $\log(-\log(1-\mu))$. Dilution assays - biology.

Models for **Proportions**: Tolerance Distributions and the Link function

Consider t_i is the tolerance level that follows a normal distribution with mean tolerance τ_i, so that

$$t_i \sim N(\tau_i, \sigma^2)$$
$$\tau_i = \beta_0' + \beta_1' x_i$$

The variable of interest is whether or not a success has occurred (the turbines develop fissures). Assume that a success if the tolerance level t_i is less than some fixed tolerance threshold T. In other words, define

$$y_i = 1 \text{ if } t_i \leq T \text{ (success); otherwise } y_i = 0.$$

Therefore, the probability of a success:

$$\mu_i = P(y_i = 1) = P(t_i \leq T) = \Phi(\frac{t_i - T}{\sigma}) \tag{72}$$

This is the probit link function.

Link function	Tolerance distribution	Distribution function
Logit	Logistic	$\mathcal{F}(y) = \exp(y)/\{1+\exp(y)\}$
Probit	Normal	$\mathcal{F}(y) = \Phi(y)$
Complementary log-log	Extreme value	$\mathcal{F}(y) = 1 - \exp\{-\exp(y)\}$
Cauchit	Cauchy	$\mathcal{F}(y) = \{\arctan(y) + 0.5\}/\pi$

Models for **Proportions**: Overdispersion

For a binomial distribution $var[y] = \mu(1-\mu)$ (μ is the probability of success) However, in practice the amount of variation in the data can exceed $\mu(1-\mu)$ (μ. This is called **overdispersion**. Underdispersion is also possbile but less common.

- Overdispersion can causes that standard errors returned by the glm are underestimated.
- Overdispersion Can be detected by conducting a goodness-of-fit test [make sure all diagnostics are good].
- Overdispersion can arise also when the trials are not independent. Clustered data.
- Modeling: Quasi-binomial where the $var[y_i] = \phi\mu(1-\mu)$; now $\phi > 0$
- Quasi-binomial model is not an EDM.
- $\hat{\beta}_{j,Quasi}$ are the same as those from binomial model but the $SE(\hat{\beta}_{j,Quasi}) = \sqrt{\phi} SE(\hat{\beta}_j)$

Remarks

- For binomial model, we need to be cautious about Hauck-Donner Effect where Wald test fails -> use LR or score tests. When the linear predictor includes factors, sometimes in practice there is a factor level for which the y_i are either all zero or all one.
- If data is ungrouped (e.g. 0 and 1) then Goodness-Of-Fit tests are not appropriate.
- If data is ungrouped (e.g. 0 and 1) use likelihood ratio and score tests.

Models for **Counts**: Poisson and Negative Binomial

The outcomes of many studies are counts y_i:

- the number of daily COVID-19 cases in Florida.
- the number of cars accident per week in Escambia county.

The Poisson distribution is an EDM where:

- Variance function: $V(\mu) = \mu$
- $\phi = 1$.
- $\theta = \log \mu$ is the link canonical function.

The model:

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{73}$$

Sometimes the link functions *identity* or *sqrt* are used.

- When the explanatory variables are all factors the data can be summarized as a contingency table and the model is often called a **log-linear model**, otherwise is a *Poisson regression model*.
- **Modeling rates** can be appropriate to make the comparison meaningful (e.g. counts/population). An *offset* term is added to the model.

Models for **Counts**: Poisson and Negative Binomial

For a Poisson distribution, $var[y] = \mu$. However, in practice the apparent variance of the data often exceeds μ. This is called **overdispersion**. Two ways to model the overdispersion:

Negative Binomial is an EDM where $Var[y] = \mu + \psi\mu^2$. The overdispersion is modeled by considering $y_i \mid \lambda_i \sim Pois(\lambda_i)$ and $\lambda_i \sim Gamma(\mu_i, \psi)$. This mixture distribution results in the *NB* distribution $y_i \sim NB(\mu_i, K = 1/psi)$, see Slide 35.
R:-> glm.nb(){MASS}.

Quasi-Poisson is not an EDM where $Var[y] = \phi\mu$.
R:-> glm(.,family="quasipoisson"){MASS}.

Models for **Counts**: Zero-Inflated Poisson and Negative Binomial

When the frequency of Zeros is larger than expected under standard discrete models then for example the Poisson model is not appropriate if the mean is larger than Zero but still you observe a mode at Zero.

- number of times in the past week that individuals report exercising
- number of times in the past week of having an alcoholic drink

The **Zero-inflated Poisson (ZIP)** model assumes:

$$y_i = \begin{cases} 0 & \text{with probability } 1 - p_i \\ Poisson(\lambda_i) & \text{with probability } p_i \end{cases} \tag{74}$$

The mean and variance:

$$E(y_i) = p_i\lambda_i \text{ and } var(y) = p_i\lambda_i(1 + (1 - p_i)\lambda_i) \tag{75}$$

The models:

$$\log(\lambda_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \text{ and } logit(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j z_{ij} \tag{76}$$

A downside of ZIP models is the larger number of parameters compared with the classical Poisson model or NB models.

Models for **Positive Continuous**: Gamma

Many applications have response variables which are continuous and positive y_i. Such variables usually have distributions that are **right skew**. The Gamma distribution is an EDM where:

- Variance function: $V(\mu) = \mu^2$
- $\theta = 1/\mu$. In practice, log is often used because it avoids the need for constraints on the linear predictor in view of $\mu > 0$.
- $\phi = 1/\alpha$. α is the shape parameter of $Gamma(\alpha, \beta)$
- ϕ is unkown and need to be estimated. Use F-tests.

The model:

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{77}$$

The time between occurrences that follow a Poisson distribution is assumed to follow Gamma.

Tips and Steps

- Purpose: Understanding or Prediction
- Criteria: Parsimony and Accuracy
- Plot and visualize the data
- Consider transformations but easily interpretable
- Fit many models
- Investigate Residuals Plots
- Data Subsetting (generalization)
- Use the appropriate performance measure to compare models: AIC, BIC, MAPE or others. For nested models: LR, ANOVA.

References


Alan Agresti.
Foundations of linear and generalized linear models.
John Wiley & Sons, 2015.


Peter K Dunn and Gordon K Smyth.
Generalized linear models with examples in R.
Springer, 2018.


Charles E McCulloch, Shayle R Searle, and John M Neuhaus.
***Generalized, Linear, and Mixed Models*, volume 651.**
John Wiley & Sons, 2011.