Week 3: Generalised Linear Models (GLM)

3.1 Generalised Linear Models definition and examples

- **Definition**
  - The idea of a generalised linear model (GLM): Nelder and Wedderburn (1972)
  - GLM is defined as a set of independent random variables $Y_1, \dots, Y_N$ each with a distribution from the **exponential family** with the following properties:
    - The distribution of each Y_i has the **canonical form** and depends on a **single parameter** θ_i (the θ_is do not all have to be the same), thus

$$f(y_i;\theta_i) = \exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)]$$

    - The distributions of all the Y_is are of the same form (e.g. all Gaussian, or all Poisson). The joint density function for independent Y_i's is

$$f(y_1,\dots,y_N;\theta_1,\dots,\theta_N) = \prod_{i=1}^{N} \exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)] \quad (3.1.1)$$

$$= \exp\left[\sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right]$$

Y_i's are indep and with the same dist

3.1 Generalised Linear Models definition and examples

- **Cont. Definition**
  - The N parameters θ_i are typically not of direct interest.
  - We are interested in a smaller set of parameters $\beta_1, \ldots, \beta_p$, where $p < N$.
    - Suppose that $\boldsymbol{E}(Y_i) = \mu_i$ is some function of θ_i.
    - For a GLM there is a **transformation** of μ_i such that

$$\eta_i = g(\mu_i) = x_i^T \beta \qquad (3.1.2)$$

link function

Linear relationship based on x_i , β

* when defining
the link function
we need it to be consistant
with the with the rang and
nature of response
varible.

where

- g is a monotone, differentiable function called the **link function**,
- x_i is a p vector of explanatory variables (or covariates) and the ith column of the design matrix $\boldsymbol{X}$

$$x_i^T = (x_{i1}, \ldots, x_{ip})$$

- β is the p vector of parameters.

  - For responses $Y_1, \ldots, Y_N$, we can write a GLM in matrix notation as

$$g[\boldsymbol{E}(y)] = \boldsymbol{X}\beta,$$

Elements of $\boldsymbol{X}$ are constants for levels of categorical explanatory variables or measured values of continuous explanatory variables

3.1 Generalised Linear Models definition and examples

* normal dist belonges to exp family

* it's in the canonical form

$a(y_i) = y_i \quad b(\mu_i) = \frac{\mu_i}{\sigma}$

- **Example: Normal linear model**
  - The best known case of a GLM is the normal linear model

$$\boldsymbol{E}(Y_i) = \mu_i = x_i^T\beta; \quad Y_i \sim N(\mu_i, \sigma^2)$$

  - The link function is the identity function $g(\mu_i) = \mu_i$. → $\mu_i = x_i^T\beta$
  - This model is usually written in the form

$$y = \boldsymbol{X}\beta + \varepsilon$$

    where ϵ is a vector of i.i.d. random variables with $\varepsilon_i \sim N(0, \sigma^2)$.
  - In this form, the linear component $\mu = \boldsymbol{X}\beta$ represents the the 'signal' and ε represents the 'noise'.
    - Multiple regression and ANOVA (analysis of variance) are of this form

3.1 Generalised Linear Models definition and examples

- **MLE for GLMs**
  - The joint distribution is

$L(\theta_1,\dots,\theta_N)$

$= \exp(\sum y_i b(\theta_i + \cdots)$

$$f(Y_1,\dots,Y_N|\theta_1,\dots,\theta_N) = \prod_{i=1}^{N} \exp[Y_i b(\theta_i) + c(\theta_i) + d(Y_i)]$$

$$= \exp\left[\sum_{i=1}^{N} Y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(Y_i)\right]$$

  - The log-likelihood for all the Y_i's is then

$\ell = \log L(\theta_1,\dots\theta_N)$ ⟶

$$\ell(\boldsymbol{\theta}; Y_1,\dots,Y_N) = \sum_{i=1}^{N} \ell_i = \sum_{i=1}^{N} Y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(Y_i).$$

  - For each Y_i, we know that

$$\mathbb{E}(Y_i) = \mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)}, \quad \mathbb{V}\text{ar}(Y_i) = \frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{[b'(\theta_i)]^3}, \quad g(\mu_i) = \mathbf{x}_i^\top \beta = \eta_i.$$

3.1 Generalised Linear Models definition and examples

- **Cont. MLE for GLMs**
  - The score function is then given by

$$U_j = \sum_{i=1}^{N} \left[\frac{(Y_i - \mu_i)}{\mathbb{V}ar(Y_i)} X_{ij} \left(\frac{d\mu_i}{d\eta_i} \right) \right] \tag{3.1.4}$$

Detailed proof

  - The variance-covariance matrix of the score is

$$\mathcal{I}_{jk} = \sum_{i=1}^{N} \frac{X_{ij} X_{ik}}{\mathbb{V}ar(Y_i)} \left(\frac{d\mu_i}{d\eta_i} \right)^2 \tag{3.1.6}$$

Detailed proof

  - How to apply the method of scoring to approximate the MLE??

Provide details

3.2 Logistic Regression

multiple logistic regression (more than 1 predictor)

- **General logistic regression** —> multinomial logistic regression (more than 2 option for Y)
  - ○ Consider a model where the outcome variables are measured on a binary scale.
    - Define a binary random variable

$P(Y_i = 1) = \pi_i = 1 - P(Y_i = 0)$

$P(Y_i = y) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$ *

$$Y = \begin{cases} 1 & \text{if the outcome is a “success”} \quad \pi \\ 0 & \text{if the outcome is a “failure”} \quad (1-\pi) \end{cases} \qquad (3.2.1)$$

i.e. Y has a Bernoulli distribution, $Y \sim B(\pi)$.

$Y \sim B(n, \pi)$

Binomial dist

n: # of repeat

π : prob. of success

  - ○ **Goal**: The goal is to relate π_i, to a set of explanatory variables $\mathbf{x}_i^\top$.
  - ○ The joint likelihood function is

$$f(\mathrm{Y}_1,\ldots,\mathrm{Y}_N|\boldsymbol{\pi}) = {}^{*}\prod_{i=1}^{N} \pi_i^{\mathrm{Y}_i}(1-\pi_i)^{1-\mathrm{Y}_i}$$

$$\prod \exp\left(Y_i \log \pi_i + (1-Y_i)\log(1-\pi_i)\right) = \exp\left[\sum_{i=1}^{N} \overset{\text{canonical form}}{\boxed{\mathrm{Y}_i}}\underbrace{\log\left(\frac{\pi_i}{1-\pi_i}\right)}_{\text{natural parameter} \rightarrow \text{canonical link function}} + \sum_{i=1}^{N}\log(1-\pi_i)\right] \qquad (3.2.3)$$

3.2 Logistic Regression

$Y_i \sim B(\pi_i) \longrightarrow E(Y_i) = \pi_i$ and $Var(Y_i) = \pi_i(1-\pi_i)$

- **Cont. General logistic regression**
  - We want to describe the probability of success with respect to some predictors:

$$g(\pi_i) = \mathbf{x}_i^\top \beta \tag{3.2.4}$$

→ 0 and 1

Note that

- The response variable is binary and not continuous
- The response variable is bounded (in $[0, 1]$)
- The variance is not constant $\mathbb{V}ar(Y_i) = \pi_i(1-\pi_i)$

Similar considerations apply to ordinal response variables.

Example and R Code

3.2 Logistic Regression

- **Example 1: Predicting the medical condition based on the symptoms**
  - Suppose there are three possible diagnoses:

$$Y = \begin{cases} 1 & \text{stroke} \\ 2 & \text{drug overdose} \\ 3 & \text{epileptic seizure} \end{cases}$$

  - Using a linear regression would assume
    - The ordering is meaningful: numbers 1, 2 and 3 are just labels!
    - The difference between ”stroke” and ”drug overdose” has the same meaning than that between ”drug overdose” and ”epileptic seizure”

3.2 Logistic Regression

- **Cont. Example 1.**
- The general logistic regression model is

natural parameter ≡ canonical link function

$$\text{logit}(\pi_i) = \underbrace{\log\left(\frac{\pi_i}{1-\pi_i}\right)} = \mathbf{x}_i^\top \beta \qquad (3.2.5)$$

where

- $\mathbf{x}_i$ is a vector of either continuous measurements or categorical variables
- β is a parameter vector.

3.2 Logistic Regression

$$\pi_i \longrightarrow 0 \implies \frac{\pi_i}{1-\pi_i} \longrightarrow 0$$

$$\pi_i \longrightarrow 1 \implies \frac{\pi_i}{1-\pi_i} \longrightarrow \infty$$

- **Cont. Example 1.**
  - $\frac{\pi_i}{1-\pi_i} \in [0, \infty)$ is an odds, indicating very low and very high probability π_i.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^\top \beta \qquad \Rightarrow \qquad \pi_i = \frac{\exp[\mathbf{x}_i^\top \beta]}{1+\exp[\mathbf{x}_i^\top \beta]}$$

  - The log-likelihood can be rewritten with respect to β and β can be estimated by maximizing $\ell(\beta; \mathbf{y}, \mathbf{x})$:

$$\ell(\beta; \mathbf{y}, \mathbf{x}) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{\exp[\mathbf{x}_i^\top \beta]}{1+\exp[\mathbf{x}_i^\top \beta]}\right) + (1 - y_i) \log\left(\frac{1}{1+\exp[\mathbf{x}_i^\top \beta]}\right)\right] \qquad (3.2.6)$$

  - The estimation process is the same if $Y_i \sim \mathrm{Bin}(n, \pi)$ (modify to consider n).
  - If the goal is prediction, one might predict

$$Y_{N+1} = 1 \qquad \text{if } \pi_{N+1} | \mathbf{x}_{N+1}^\top > 0.5.$$

Other thresholds could also be used, e.g., to be conservative, set the threshold to 0.1.

3.2 Logistic Regression

- **Example 2: Analysis of trade union dataset**

  Example and R Code

- **Prediction**
  - Once the coefficients have been estimated, predictions are obtained by using those estimates with the desired level of predictors.

  Example and R Code

3.2 Logistic Regression

- **Goodness of fit**
  - In a linear model, residual plots are useful in exhibiting violations of model assumptions (e.g. independence, homoscedasticity).
  - In a GLM, we would like to assign a residual e_i to each observation which measures the discrepancy between Y_i and the value predicted by the fitted model.
  - There are two main difficulties associated with generalised linear models:
    - The model variances depend on the expectations; Binomial $\mu_i = n_i p_i$, $\sigma^2 = n_i p_i (1-p_i)$
    - It is not obvious that data and fitted values should be compared on the original scale of the responses. Poisson

$\mu_i = \lambda_i$

$\sigma^2 = \lambda_i$

link function $g(\mu_i) = x_i^T \beta$

logistic regression $y_i = 0$ or 1 while fitted values $\underbrace{\hat{\pi}_i \in (0,1)}_{\hat{y}_i}$

set a threshold

3.2 Logistic Regression

- **Goodness of fit: Pearson chi-squared statistic**
  - **Pearson residuals**: the difference between observed and fitted values, divide by an estimate of the standard deviation of the observed values.
  - For $Y_i \sim \mathrm{Bin}(n_i, \pi_i)$, the Pearson residuals are

$$P_i = \frac{(y_i - n_i\hat{\pi}_i)}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}, \qquad i = 1, \ldots, N.$$

  - Instead of maximising the likelihood, estimate the parameters by minimising the weighted sum of squares

$$S_w = \sum_{i=1}^{N} \frac{(y_i - \mathbb{E}(\mathrm{Y}_i))^2}{\mathbb{V}\mathrm{ar}(\mathrm{Y}_i)} = \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}$$

  - If o_i is the observed and e_i is the expected frequencies, then **Pearson chi-squared statistic** is

$$P^2 = \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{e_i}.$$

3.2 Logistic Regression

- **Cont. Goodness of fit: Pearson chi-squared statistic**
  - The reason for the equivalence is

expected values

$$\sum \frac{(o_i - e_i)^2}{e_i}$$

$$P^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i} + \sum_{i=1}^{N} \frac{[(n_i - y_i) - n_i(1-\pi_i)]^2}{n_i(1-\pi_i)}$$

$$= \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}(1 - \pi_i + \pi_i) = S_w$$

  - P^2 is evaluated at the estimated expected frequencies; i.e., $P^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$

Frequencies for N Binomial distributions.

observed

	Subgroups			
	1	2	...	N
Successes	Y_1	Y_2	...	Y_N
Failures	$n_1 - Y_1$	$n_2 - Y_2$	...	$n_N - Y_N$
Totals	n_1	n_2	...	n_N

3.2 Logistic Regression

- **Goodness of fit: Deviance**
  - The deviance for the logistic model is

$$D = 2\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right] \qquad (3.2.7)$$

Proof

  - The deviance is asymptotically equivalent to the Pearson chi-squared statistic evaluated at the estimated expected frequencies. (Proof)
  - Under the null hypothesis (H_0), the asymptotic distribution of D is

$$D \sim \chi^2(N-p) \qquad \Rightarrow \qquad P^2 \sim \chi^2(N-p) \qquad (3.2.8)$$

  - The adequacy of the approximation depends on how well D or P^2 are χ^2-distributed.
  - There is some evidence that P^2 is better than D, however both of them are influenced by small frequencies. This is typical of continuous covariates.

Example and R Code

3.2 Logistic Regression

- **Goodness of fit: Hosmer-Lemeshow Statistic**
  - A possible solution is to group observations, based on their predicted probabilities.

    $g \rightarrow$ less than 10
  - Each group has approximately equal numbers of observations.
  - The Pearson chi-squared statistic is computed on the contingency table obtained by grouping observations.

    ↳ 2 rows ↗ success ↘ failure

    g columns
  - This statistic is called Hosmer-Lemeshow statistic.

    ↳ $\simeq \chi^2_{g-2}$

R Code

3.2 Logistic Regression

- **Likelihood ratio, Pseudo R^2, AIC and BIC**
  - **Likelihood ratio χ^2 statistic**
    - Compare the log-likelihood of the fitted model with the log-likelihood of the minimal model, in which all π_i are equal and $\tilde{\pi} = \sum_{i=1}^{N} y_i / \sum_{i=1}^{N} n_i$.
    - The statistic is defined as

$$C = 2[\ell(\hat{\pi}; \mathbf{y}) - \ell(\tilde{\pi}; \mathbf{y})]$$

$$= 2\sum_{i=1}^{N} \left[y_i \log\left(\frac{\hat{y}_i}{n_i\tilde{\pi}}\right) + (n_i - y_i)\log\left(\frac{n_i - \hat{y}_i}{n_i - n_i\tilde{\pi}}\right)\right] \sim \chi^2(p-1)$$

R Code

3.2 Logistic Regression

- **Pseudo-R^2**
  - Analogously to the multiple LR, the likelihood ratio statistic can be normalised

$$\text{pseudo-}R^2 = \frac{\ell(\tilde{\pi};\mathbf{y}) - \ell(\hat{\pi};\mathbf{y})}{\ell(\tilde{\pi};\mathbf{y})} \qquad (3.2.9)$$

  - It represents the proportional improvement in the log-likelihood function due to the terms in the model of interest, compared with the minimal model.
  - As for R^2, the distribution of the pseudo-R^2 cannot be determined, and it increases as the number of predictors increases. Therefore, several adjustments have been proposed.

R Code

3.2 Logistic Regression

- **Cont. Likelihood ratio, Pseudo R^2, AIC and BIC**
  - **AIC and BIC**
    - The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are very popular goodness of fit statistics based on the log-likelihood, with an adjustment for the number of parameters, p, and the sample size, N.

$$\mathrm{AIC} = -2\ell(\hat{\boldsymbol{\pi}}; \mathbf{y}) + 2p \qquad (3.2.10)$$

$$\mathrm{BIC} = -2\ell(\hat{\boldsymbol{\pi}}; \mathbf{y}) + p \times \log N \qquad (3.2.11)$$

    - **Remark**: all these statistics (except the pseudo-R^2) summarise how well a particular model fits the data: a small value (or a large p-value) indicates that the model fits well.

R Code

3.2 Logistic Regression

- **Residuals**
  - For $Y_i \sim \mathrm{Bin}(n_i, \pi_i)$, the **Pearson Residuals** are

$$P_i = \frac{(Y_i - n_i\hat{\pi}_i)}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}, \quad i = 1, \ldots, N$$

    which can be **standardised** by dividing by the **leverage** h_{ii}

$$e_{iP} = \frac{P_i}{\sqrt{1-h_{ii}}}$$

  - Notice that $\sum_{i=1}^{N} P_i^2 = P^2$
  - The **Deviance Residuals** are defined as

$$d_i = \mathrm{sign}(Y_i - n_i\hat{\pi}_i)\left\{2\left[Y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - Y_i}{n_i - n_i\hat{\pi}_i}\right)\right]\right\}^{1/2}$$

    (the sign term makes sure that the signs of d_i and P_i match).
  - Note that $\sum_{i=1}^{N} d_i^2 = D$, the deviance.

3.2 Logistic Regression

- **Cont. Residuals**
  - The residuals can be used in the usual way: they should be
    - Plotted against each continuous explanatory variable (check linearity assumption)
    - Plotted against other possible explanatory variables not included in the model
    - Plotted in the order of the measurements to check for correlation
    - Through normality plots → standardised residuals should be normal &
  - For GLM, residual plots are less informative than for multiple LR, therefore check all the other goodness-of-fit statistics.
  - Plot residuals against the values of the linear predictors in GLM to look for patterns in the residuals (related to the mean).
  - Sometimes, it can be hard to see patterns in residual plots for GLM.
    - In logistic regression, with binary responses, the residuals can only take on two possible values (depending on whether the response is zero or one) and when residuals are plotted against linear predictor values, all points lie on one of the two smooth curves.
  - Superimpose a scatterplot smoother on the residual plot to identify any trends.

Example and R Code

3.3 Poisson Regression

- **Poisson regression and Log-Linear regression**
  - Poisson distribution is used for count data.
  - If a random variable Y is Poisson distributed then it has probability distribution

$$\exp\{y \overbrace{\log\mu}^{\text{natural parameter}} - \mu - \log y!\}$$

canonical form

$$E(Y) = Var(Y) = \mu$$

$$f(y) = \frac{\mu^y e^{-\mu}}{y!} \qquad y = 0, 1, \ldots \qquad (3.3.1)$$

where μ is a parameter such that $\mathbb{E}(Y) = \mu$ and which is often called ”rate”.

Example

The number of tropical cyclones crossing the North Queensland coast can be represented as a Poisson random variable.
μ : the rate of tropical cyclones crossing the North Queensland coast in the cyclone season, from November to April.

3.3 Poisson Regression

effect of different marketing advertisments of sales

- the number of people exposed to diff types of advertisment are constant

response → sale   predictor → type of advertisement

- **Cont. Poisson regression and Log-Linear regression**
  - The effect of explanatory variables on the response Y is modelled through the parameter μ.
    - **Poisson regression**: the events relate to varying amounts of exposure which need to be taken into account when modelling the rate of events (explanatory variables are usually continuous or categorical)
    - **Log-linear regression**: exposure is constant (explanatory variables are usually categorical)

rate of hospital admissions for a specific disease across diff regions.

response → # of hospital admissions

predictors → average age in the region, income level

Exposure → population in each area → offset

3.3 Poisson Regression

- **Poisson regression**
  - Let $Y_1, \ldots, Y_N$ be independent random variables, with Y_i denoting the number of events observed from exposure n_i for the i-th covariate pattern.
  - The expected value of Y_i is

$$\mathbb{E}(Y_i) = \mu_i = n_i\theta_i. \tag{3.3.2}$$

Parameter θ_i depends on a set of explanatory variables $\mathbf{x}_i$ and is modelled as

$$\theta_i = e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \tag{3.3.3}$$

  - The natural link function is the logarithmic function

$$\log \mu_i = \log n_i + \mathbf{x}_i^\top \boldsymbol{\beta} \tag{3.3.4}$$

  - The term $\log n_i$ is called the **offset**, a known constant.

3.3 Poisson Regression

- **Cont. Poisson regression**
  - **Interpretation in terms of rate ratio (RR):** Suppose we have a dummy variable

$$x_j = \begin{cases} 0 & \text{if factor is absent} \\ 1 & \text{if factor is present} \end{cases}$$

Then, if the other explanatory variables stay the same, RR for presence versus absence is

$$\mathrm{RR} = \frac{\mathbb{E}(Y_i|present)}{\mathbb{E}(Y_i|absent)} = e^{\beta_j} \tag{3.3.5}$$

  - Similarly, for a continuous explanatory variable x_l, e^{β_l} represents the multiplicative effect on the rate μ.

$$x_l + 1 \longrightarrow \mu \times e^{\beta_l}$$

3.3 Poisson Regression

- **Residuals for the Poisson model**
  - Once the regression coefficients are estimated through $\hat{\beta}$, the fitted values are given by

$$\hat{\mathrm{Y}}_i = \hat{\mu}_i = n_i e^{\mathbf{x}_i^\top \hat{\beta}}, \qquad i = 1, \ldots, N \tag{3.3.6}$$

  - Similarly to LR, we can call these fitted values e_i since they estimate the expected values $\mathbb{E}(\mathrm{Y}_i) = \mu_i$.

    expected values
  - Using the fact that $\underline{\mathbb{V}\mathrm{ar}(\mathrm{Y}_i)} = \underline{\mathbb{E}(\mathrm{Y}_i)}$, the **Pearson Residuals** are

$$\frac{Y_i - E(Y_i)}{\sqrt{Var(Y_i)}}$$

$$P_i = \frac{o_i - e_i}{\sqrt{e_i}}$$

where o_i (or Y_i) denotes the observed count and e_i (or $\hat{\mathrm{Y}}_i$) the expected count.

- The Pearson residuals are used to compute the Pearson chi-squared goodness of fit statistic

$$P^2 = \sum_{i=1}^{N} P_i^2 = \sum_{i=1}^{N} \frac{(o_i - e_i)^2}{e_i} \tag{3.3.7}$$

3.3 Poisson Regression

- **Cont. Residuals for the Poisson model**
  - The deviance residuals are

$$d_i = \operatorname{sign}(o_i - e_i)\sqrt{2[o_i \log(o_i/e_i) - (o_i - e_i)]}, \qquad i = 1, \dots, N$$

glm

Again, $D = \sum d_i^2$ is the deviance and

$$D = 2\sum_{i=1}^{N}[o_i \log(o_i/e_i) - \cancel{(o_i - e_i)}] \qquad (3.3.8)$$

Since in many cases $\sum_{i=1}^{N} o_i = \sum_{i=1}^{N} e_i$, we have $D = 2\sum_{i=1}^{N}[o_i \log(o_i/e_i)]$. Use a Taylor expansion of $o_i \log(o_i/e_i)$:

$$o_i \log(o_i/e_i) = (o_i - e_i) + \boxed{\frac{1}{2}\frac{(o_i - e_i)^2}{e_i}} + \cdots,$$

$$\Rightarrow D = 2\sum_{i=1}^{N}\left[(o_i - e_i) + \frac{1}{2}\frac{(o_i - e_i)^2}{e_i} - (o_i - e_i)\right] = \sum_{i=1}^{N}\left[\frac{(o_i - e_i)^2}{e_i}\right] = P^2$$

3.3 Poisson Regression

- **Cont. Residuals for the Poisson model**
  - An important aspect of D and P^2 is that they depend on the fitted values and the observations, and do not depend on any nuisance parameters (like σ^2 for the Normal).

Example (Poisson regression)

Consider the artificial dataset `poisson` from the `dobson` package in `R` and which represents counts `y` observed at various values of a covariate `x`.

code

3.3 Poisson Regression

Example (Cont. Poisson regression)

Assumption: $Y_i \sim Poisson(\mu_i)$

- **Why this assumption**?? variability increases with Y.

If $Y_i \sim Poisson(\mu_i)$, then $\mathbb{E}(Y_i) = \mathbb{V}ar(Y_i)$. Let us model the relationship between Y_i and X_i by the straight line

$$\mathbb{E}(Y_i) = \mu_i = \beta_1 + \beta_2 X_i = \mathbf{x}_i^\top \beta$$

for $i = 1, \ldots, N$, where

$$\beta = (\beta_1, \beta_2)^\top, \quad \text{and} \quad \mathbf{x}_i = (1, X_i)^\top.$$

3.3 Poisson Regression

general : $\eta_i = g(\mu_i)$

section 2.2

subsection 1.1

- **Modelling with the identity link function**
  - Take the link function to be the identity function:

$\eta_i = \mu_i = x_i^T\beta \longrightarrow$ link function is identity

$$g(\mu_i) = \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \eta_i \qquad \Rightarrow \qquad d\eta_i/d\mu_i = 1.$$

In this case we have

$$w_{ii} = \frac{1}{\mathbb{V}\mathrm{ar}(Y_i)} = \frac{1}{\beta_1 + \beta_2 X_i}.$$

$I = X^T W X$

$I^{(m-1)}\hat{\beta}^{(m-1)} + U^{(m-1)}$

$= X^T W z$

Using an estimate $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)^\top$ for β, we obtain

$$Z_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} + (Y_i - \overset{x_i^T\hat{\beta}}{\widetilde{\mu_i}})\left(\frac{d\eta_i}{d\mu_i}\right)$$
$$= \cancel{\hat{\beta}_1} + \cancel{\hat{\beta}_2 x_i} + (Y_i \cancel{- \hat{\beta}_1} \cancel{- \hat{\beta}_2 X_i})$$
$$= Y_i$$

3.3 Poisson Regression

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \quad W = \mathrm{diag}\left(\frac{1}{\hat{\beta}_1 + \hat{\beta}_2 x_i}\right)$$

- **Cont. Modelling with the identity link function**

  Additionally,

$$\mathcal{I} = \mathbf{X}^\top \mathbf{W} \mathbf{X} = \begin{pmatrix} \sum_{i=1}^N \frac{1}{\hat{\beta}_1+\hat{\beta}_2 x_i} & \sum_{i=1}^N \frac{x_i}{\hat{\beta}_1+\hat{\beta}_2 x_i} \\ \sum_{i=1}^N \frac{x_i}{\hat{\beta}_1+\hat{\beta}_2 x_i} & \sum_{i=1}^N \frac{x_i^2}{\hat{\beta}_1+\hat{\beta}_2 x_i} \end{pmatrix}.$$

  and

$$\mathbf{X}^\top \mathbf{W} \mathbf{z} = \begin{pmatrix} \sum_{i=1}^N \frac{y_i}{\hat{\beta}_1+\hat{\beta}_2 x_i} \\ \sum_{i=1}^N \frac{x_i y_i}{\hat{\beta}_1+\hat{\beta}_2 x_i} \end{pmatrix}.$$

  The maximum likelihood estimates are obtained iteratively from the equations

$$(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{(m-1)} \hat{\beta}^{(m)} = \mathbf{X}^\top \mathbf{W} \mathbf{z}^{(m-1)}.$$

3.3 Poisson Regression

Example (Cont. Poisson regression)

- Code:
  1. Initial points to start the iteration in finding MLE
  2. Obtaining MLE
  3. Find the 95% confidence intervals
  4. Use `glm` function in `R`

3.3 Poisson Regression

- **Modelling with the canonical link function**

Example (Cont. Poisson regression)

- Code:
  1. Use the canonical link function
  2. The rate ratio of increasing the predictor of 1 unit
  3. The Pearson residuals
  4. The Pearson chi-squared goodness of fit statistic
  5. The deviance statistic
  6. The pseudo R^2

3.4 Log-linear regression

- Log-linear models:

$$\log \mathbb{E}(Y_i) = c + \mathbf{x}_i^\top \beta \tag{3.4.1}$$

We will analyse

- ○ Analyse the introduction of interaction terms.
- ○ Analogous of the ANOVA for log-linear models.

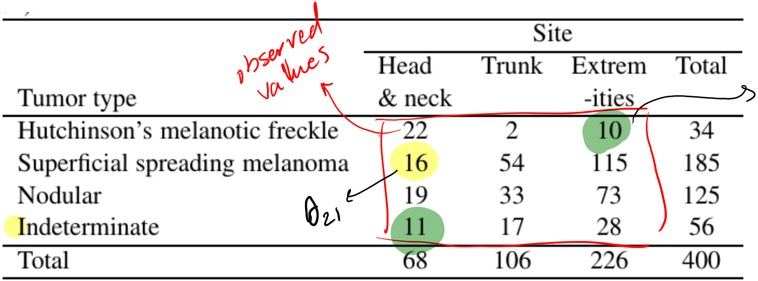
3.4 Log-linear regression

Example (Melanoma Dastset)

- Cross-sectional study of patients some skin cancer
- $N = 400$ → constant
- Information about the site and the histological type of the tumour.

categorical predictors

- contingency table



	Site			
Tumor type	Head & neck	Trunk	Extrem -ities	Total
Hutchinson's melanotic freckle	22	2	10	34
Superficial spreading melanoma	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

observed values

θ_{21}

estimated freq in saturated model

3.4 Log-linear regression

θ_{jk}: probability of being in cell (j, k)

- In case of no association: → site and type are indep

$$P(A\cap B)=P(A)P(B) \leftarrow \theta_{jk} = \theta_{j.}\theta_{.k} \qquad j = 1, \dots, J \quad \text{and} \quad k = 1, \dots, K \tag{3.4.2}$$

- In the case of independence → $E(Y)_{jk} = n\,\theta_{jk}$ → $\theta_{j.}\,\theta_{.k}$

$$\log \mathbb{E}(\mathrm{Y})_{jk} = \log n + \log \theta_{j.} + \log \theta_{.k} \tag{3.4.3}$$

which can be compared with the dependent model, i.e.

$$\log \mathbb{E}(\mathrm{Y})_{jk} = \log n + \log \theta_{jk} \tag{3.4.4}$$

3.4 Log-linear regression

- Analogously to ANOVA, introduce the factors relative to the single predictors/factors

$$\log \mathbb{E}(\mathrm{Y})_{jk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}, \qquad \text{Saturated Model} \tag{3.4.5}$$

  where $(\alpha\beta)_{jk}$ is a coefficient relative to the interaction term.

- To test for independence, we can compare (3.4.5) with

$$\log \mathbb{E}(\mathrm{Y})_{jk} = \mu + \alpha_j + \beta_k, \qquad \text{Additive Model}$$

  or, since $\log(n)$ is in all the models,

$$\log \mathbb{E}(Y)_{jk} = \mu, \qquad \text{minimal Model}$$

3.4 Log-linear regression

- The specification of log-linear models is hierarchical: if the higher-order term (interaction) is included in the model, all the lower-order terms are included as well. $(\alpha\beta)_{jk} \longrightarrow \alpha_j$ and β_k
- **Warning**: In many cases, log-linear models have many parameters: constraints may be needed!
- While several distributions can be used, Poisson distributions can be assumed. Therefore, all standard methods for GLM can be applied (weighted least squares, goodness-of-fit statistics like P^2 and D, Pearson and deviance residuals).

3.4 Log-linear regression

- **Code**
  - The saturated model
  - The model with no interaction terms
  - The minimal model

3.4 Log-linear regression

base levels

- For the reference category `type:Hutchinson’s melanotic freckle` on `site:extremities` the expected frequencies are
  - minimal model: $e^{3.507} = 33.33$ $\longrightarrow e^{\hat{\beta}_{1,\text{min}}}$
  - additive model: $e^{2.9554} = 19.21$ $\longrightarrow e^{\hat{\beta}_{1,\text{addi}}}$
  - saturated model: $e^{2.3026} = 10.00$ $\longrightarrow e^{\hat{\beta}_{1,\text{satur}}}$

  Note: the expected frequencies for the saturated model correspond to the observed frequencies.
- For `type:indeterminate` tumours on `site:head-neck` the expected frequencies are
  - minimal model: $e^{3.507} = 33.33$ $\longrightarrow e^{\hat{\beta}_{1,\text{minal}}}$
  - additive model: $e^{2.9554-1.2010+0.499} = 9.520049$ $\longrightarrow e^{\hat{\mu} + \hat{\alpha}_{\text{indet}} + \hat{\beta}_{\text{head-neck}}}$
  - saturated model: $e^{2.3026+0.7885+1.0296-1.7228} = 11.000$ = observed value

  Again the expected frequencies for the saturated model correspond to the observed frequencies.

3.4 Log-linear regression

- For `type:nodular` tumours on `site:trunk` the expected frequencies are
  - minimal model: $e^{3.507} = 33.33$
  - additive model: $e^{2.9554-0.7571+1.3020} = e^{3.5003} = 33.12$
  - saturated model: $e^{2.3026-1.6094+1.9879+0.8155} = e^{3.4966} = 33.00$

- saturated model fit the data accurately