SIMULATION FROM GLM

12/03/25

Two objectives:

1. Learn the idea behind simulation

2. Recall the theory behind GLMs

We are going to consider 2 models:

1. Poisson regression (count response)

2. Logistic regression (classification model, binary response)

EX 1 POISSON REGRESSION

Y response (count) $\quad \underline{X} = X_1, \dots, X_p$

$Y_i = Y \mid \underline{X} = \underline{x}_i \,, \quad i = 1, \dots, n$

$$
\begin{array}{c|ccc|c}
 & 1 & \cdots & p & \\
\hline
1 & x_{11} & \cdots & x_{1p} & y_1 \\
\vdots & \vdots & & & \vdots \\
 & & \underline{x}_i & & y_i \\
\vdots & \vdots & & & \vdots \\
n & x_{n1} & \cdots & x_{np} & y_n
\end{array}
$$

$y_1, \dots, y_n$ observations of Y_i
(independent but not identically distributed)

$\mu_i = E[Y_i] \in [0, \infty)$

$\eta_i = \underline{x}_i^t \underline{\beta} \in (-\infty, \infty)$

$\log(\mu_i) = \eta_i$ canonical link (g)

$\iff \mu_i = e^{\eta_i}$

$Y_i \sim \text{Poisson}(\mu_i) = \text{Poisson}(e^{\eta_i}) = \text{Poisson}(e^{\underline{x}_i^t \underline{\beta}})$

$$\boxed{Y_i \sim \underbrace{\text{Poisson}}_{(f)}\big(\underbrace{e^{\underline{x}_i^t \underline{\beta}}}_{g^{-1}}\big)}$$

SIMULATION

$p = 1 \,, \quad n = 100$

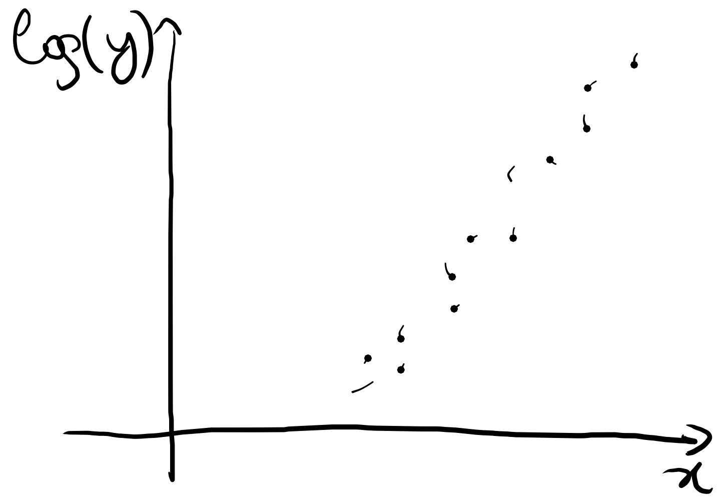
$$
\begin{array}{cc}
x & y \\
\hline
\vdots & \vdots
\end{array}
$$

1. Generate $x_1, \dots, x_{100}$ from any distribution

2. $\eta_i = \beta_0 + \beta_1 x_i$

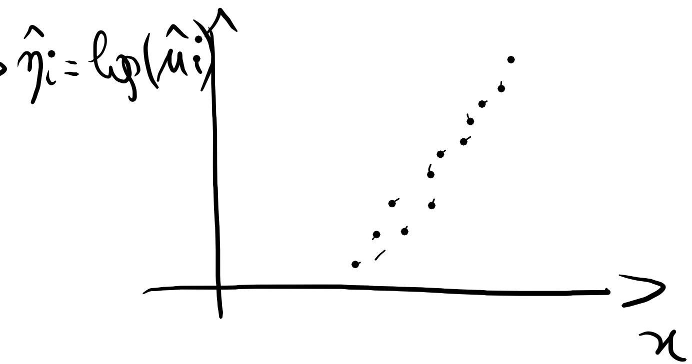
   Fix $\beta_0 = 1, \; \beta_1 = 2$

3. Sample y_i from $\text{Poisson}(e^{\eta_i}), \; i = 1, \dots, 100$



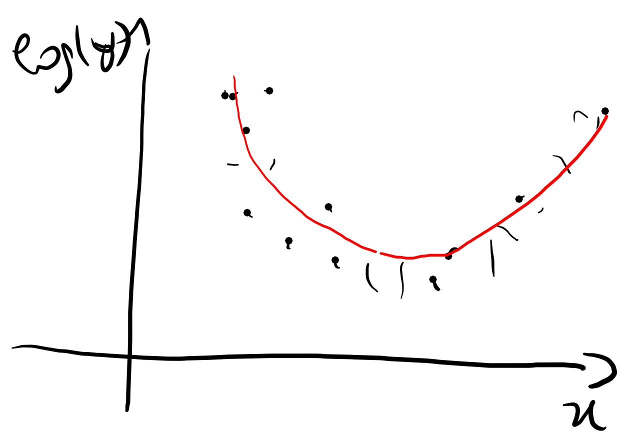
Non-parametric estimate of μ_i is

$$\hat{\mu}_i = y_i$$



$\log(\mu_i) = \beta_0 + \beta_1 x_i$ is a good idea!

If instead:



This would suggest the inclusion of x^2 in the linear predictor

RESIDUALS

In linear models:

$$y_i = \underset{\underset{x_i^t \beta}{\shortparallel}}{\eta_i} + \varepsilon_i \iff y_i \sim \text{Normal}(\underline{x}_i^t \underline{\beta}, \sigma^2)$$

(signal + noise) (modern definition)

$\downarrow$

GLMs

$$e_i = y_i - \hat{\eta}_i$$
$$= y_i - \underline{x}_i^t \hat{\underline{\beta}}$$

For generalized linear models, the two main definitions of residuals are:

1. Pearson residuals

2. Deviance residuals

PEARSON RESIDUALS

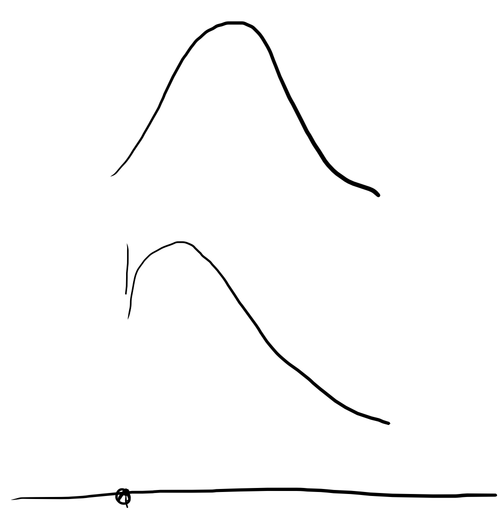
$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var}(y_i)/\hat{\phi}}}$$

For Poisson:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \quad , i=1,..,n$$

$Var(Y_i) = E[Y_i] = \mu_i$

$\phi = 1$



→ Gaussian distribution for r_i^P is an approximation

→ The larger μ_i is, the better the approximation

DEVIANCE

The deviance of observation i is defined by:

$\hat{\mu}_i = e^{x_i^t \hat{\beta}}$

$$d_i = 2\left(l_i(y_i ; y_i) - l_i(\hat{\mu}_i ; y_i) \right) , \quad i = 1, \ldots, n$$

$\hat{\mu}_i = y_i$ log-likelihood

$$d_i \geq 0$$

$$\Rightarrow \quad D = \sum_{i=1}^{n} d_i \quad \text{deviance}$$

$$\Rightarrow \quad r_i^D = \sqrt{d_i}\, \operatorname{sign}(y_i - \hat{\mu}_i) = \begin{cases} \sqrt{d_i} & y_i \geq \hat{\mu}_i \\ -\sqrt{d_i} & y_i < \hat{\mu}_i \end{cases}$$