■ Logistic regression.
■ Generalized linear models.
■ Deviance.
Generalized linear models

- All models we have seen so far deal with continuous outcome variables with no restriction on their expectations, and (most) have assumed that mean and variance are unrelated (i.e. variance is constant).
- Many outcomes of interest do not satisfy this.
- Examples: binary outcomes, Poisson count outcomes.
- A Generalized Linear Model (GLM) is a model with two ingredients: a link function and a variance function.
  - The link relates the means of the observations to predictors: linearization
  - The variance function relates the means to the variances.
A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic.

In a pilot follow-up study, 50 clients were randomly selected and asked whether they actually received a flu shot.

In addition, data were collected on their age and their health awareness.

Here is the data.
Binary outcomes

- Suppose outcome $Y_i$ is a 0-1 random variable, then
  \[ \mu_i = \mathbb{E}(Y_i) = \pi_i. \]

- Variance function
  \[ \text{Var}(Y_i) = \pi_i(1 - \pi_i) = \mu_i(1 - \mu_i) \]
  
  *Variance is related to mean!*

- A convenient way to model the dependence of $Y_i$ on covariates $X_{i1}, \ldots, X_{ip-1}$ is through the logit transform.
  \[ \text{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij} \]
Logit transform

- Logit transform:

\[ \logit(\pi) = \log \left( \frac{\pi}{1 - \pi} \right) \in (-\infty, +\infty) \]

- Inverse:

\[ \logit^{-1}(x) = \frac{e^x}{1 + e^x} \in (0, 1). \]

- Derivative:

\[ \frac{d}{d\pi} \logit(\pi) = \frac{1}{\pi(1 - \pi)} = \frac{1}{V(\pi)}. \]

**Note**: special relation between derivatives and variance function – more on this next lecture.
Binary regression

- Logistic regression model:

\[
\text{logit}(\mathbb{E}(Y_i)) = \text{logit}(\pi_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}
\]

- Probit regression model:

\[
\Phi^{-1} (\mathbb{E}(Y_i)) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}
\]

where \(\Phi\) is CDF of \(N(0, 1)\), i.e. \(\Phi(t) = \text{pnorm}(t)\).

- In each case, \(\text{Var}(Y_i) = \pi_i (1 - \pi_i)\) but the regression model is different.

- Here is an example
Link functions: binary regression

![Graph showing logit and probit functions](image-url)
Link function inverses: binary regression

The figure illustrates the relationship between the link function inverses and the probability $\pi$. The red curve represents the logit.inverse function, while the green curve represents the probit.inverse function. The x-axis represents the probability $\pi$, ranging from -4 to 4, and the y-axis represents the function value $f(\pi)$, ranging from 0 to 1. The curves show how the inverse link functions transform the probability space.
Odds ratios & logistic regression

For any event $A$ and any probability $\mathbb{P}$

$$ODDS(A) = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}.$$  

In the logistic regression model with outcome $Y$

$$\frac{ODDS(Y = 1 | \ldots, X_j = x_j + 1, \ldots)}{ODDS(Y = 1 | \ldots, X_j = x_j, \ldots)} = e^{\beta_j}$$

is the (multiplicative) change in odds if variable $X_j$ increases by 1: $e^{\beta_j}$ is known as the ODDS RATIO for $X_j$.

If $X_j \in \{0, 1\}$ is dichotomous, and $\mathbb{P}(Y = 1 | \ldots)$ is small (rare event hypothesis) then group with $X_j = 1$ are approximately $e^{\beta_j}$ more likely to have event, all other parameters being the same.
Link & variance fns. of a GLM

- If 
  \[ \eta_i = g(\mathbb{E}(Y_i)) = g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} \]
  then \( g \) is called the link function for the model.

- If 
  \[ \text{Var}(Y_i) = \phi \cdot V(\mathbb{E}(Y_i)) = \phi \cdot V(\mu_i) \]
  for \( \phi > 0 \) and some function \( V \), then \( V \) is the called variance function for the model.

- “Canonical” reference: Generalized Linear Models, McCullagh and Nelder.
For a logistic model,

\[ g(\mu) = \logit(\mu), \quad V(\mu) = \mu(1 - \mu). \]

For a probit model,

\[ g(\mu) = \Phi^{-1}(\mu), \quad V(\mu) = \mu(1 - \mu). \]
Fitting a binary regression GLM: IRLS

- **Algorithm**
  1. Initialize: set $\hat{\mu}_i = 0.999$ or 0.001 depending on whether $Y_i = 1$ or 0.
  2. Compute $Z_i = g(\hat{\mu}_i) + g'(\hat{\mu}_i)(Y_i - \hat{\mu}_i)$.
  3. Use weights $W_i^{-1} = g'(\hat{\mu}_i)^2 V(\hat{\mu}_i)$ to regress $Z$ onto $X$’s to get $\hat{\beta}$ using WLS.
  4. Compute $\hat{\mu}_i = g^{-1}(\hat{\beta}_0 + \sum_{j=1}^{p} X_{ij}\hat{\beta}_j)$.
  5. Repeat steps 2-4 until convergence.

- **Approximate distribution**
  $$\hat{\beta} \sim N(\beta, \phi(X^tWX)^{-1}).$$

- If $\phi$ has to be estimated, a simple choice is Pearson’s $X^2$:
  $$\hat{\phi} = \frac{1}{n - p} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$
Other common examples of GLMs

- Standard multiple linear regression: \( g(\mu) = \mu, \text{Var}(\mu) = 1 \)
- Linear regression with variance tied to mean, for example:
  \( g(\mu) = \mu, \text{Var}(\mu) = \mu^2. \)
- Poisson log-linear models: \( g(\mu) = \log(\mu), \text{Var}(\mu) = \mu. \)
Deviance

- Instead of least squares, models are fit on the basis of scaled deviance, analogous to SSE when errors are not Gaussian.
- This is not completely clear from the IRLS algorithm, but it is very close to the Newton-Raphson algorithm. Algorithm often called *Fisher scoring.*

\[
DEV(\mu, Y) = -2 \log L(\mu, Y) + -2 \log L(Y, Y)
\]

where \( \mu \) is a location estimator for \( Y \) (usually in an exponential family – more next lecture).

- If \( Y \) is Gaussian with independent \( N(\mu_i, \sigma^2) \) entries

\[
DEV(\mu, Y) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu_i)^2
\]
Binary deviance

- If \( Y \) is a vector of independent 0-1 random variables

\[
DEV(\mu, Y) = -2 \left( \sum_{i=1}^{n} Y_i \log \mu_i + (1 - Y_i) \log (1 - \mu_i) \right)
\]

- Uses the facts

\[
\lim_{\mu_i \uparrow 1} -2(Y_i \log \mu_i + (1 - Y_i) \log(1 - \mu_i)) = \begin{cases} 0 & Y_i = 1 \\ \infty & Y_i = 0 \end{cases}
\]

\[
\lim_{\mu_i \downarrow 0} -2(Y_i \log \mu_i + (1 - Y_i) \log(1 - \mu_i)) = \begin{cases} 0 & Y_i = 0 \\ \infty & Y_i = 1 \end{cases}
\]
Partial deviance tests

- As in multiple regression to test that some subset $H_0 : \beta_{i_1} = \cdots = \beta_{i_l} = 0$ we fit a full and a reduced model.
- Asymptotic theory tell us that

$$DEV \chi^2 = \frac{DEV(R) - DEV(F)}{\widehat{\phi}} \sim \chi^2_{df_R - df_F}.$$

If deviance is unavailable, Pearson’s $X^2$ is substituted, $\widehat{\phi}$ is analogous to $\sigma^2$.
- Reject $H_0$ if $DEV \chi^2$ is larger than $\chi^2_{df_R - df_F, 1 - \alpha}$.
Wald $\chi^2$ tests

- Test can also be done using a Wald $\chi^2$ which does not fit a full and reduced model.
- Wald $\chi^2$ to test $C\beta = 0$:
  
  $$WALD\chi^2 = C\hat{\beta}(\phi \cdot C(X^tW^{-1}X)^{-1}C^t)^{-1}(C\hat{\beta})^t.$$ 

- Reject $H_0 : C\beta = 0$ if $WALD\chi^2$ is larger than $\chi^2_{\text{#rows}C,1-\alpha}$ (assuming $C$ is full rank).