

# ISI-BUDS

## Generalized Linear Models

Babak Shahbaba, PhD  
University of California at Irvine (UCI)

# Linear models

- Linear models have been extensively used in practice.
- They include a large class of models such as ANOVA and linear regression.
- They owe their popularity mostly to the fact that they are **easy to fit and interpret**.
- We use these models to capture the relationship between the **response** variable,  $y$ , and a set of **explanatory** variables (predictors, covariates, ...),  $x$ .
- What does it mean for two random variables to be related?
- When we talk about relationship between  $y$  and  $x$ , we usually think about the change in the **conditional distribution** of  $y$  given  $x$ , i.e.,  $P(y|x)$ , as  $x$  changes.

# Relationship

- Regression models are based on the assumption that the only change in the conditional distribution we are interested in is the change in the **expectation** of the distribution,  $E(y|x)$  (note that this by itself imposes limitations on the type of relationships we can detect).
- In general, this means  $E(y|x) = g(x)$ , and the relationship between  $x$  and  $y$  exists if  $g(x)$  is not a constant function.
- In this setting,  $g(x)$  also defines the type of relationship between  $x$  and  $y$ .

# Linear regression models

- For linear regression models,  $g(x)$  is a linear function in terms of model parameters,  $\beta$ .
- Recall that a function  $f : \mathcal{R}^n \rightarrow \mathcal{R}^m$  is **linear** if
  - $f(z + t) = f(z) + f(t)$ ,  $\forall z, t \in \mathcal{R}^n$
  - $f(az) = af(z)$ ,  $\forall z \in \mathcal{R}^n, \forall a \in \mathcal{R}$
- The function  $g(x)$  has the following general form:

$$g(x) = x\beta$$

where  $x$  is a  $n \times (p + 1)$  matrix (the first column is the constant 1, and the remaining  $p$  columns are the observed value of  $p$  explanatory variables)

- $\beta$  is a  $(p + 1)$ -vector of parameters. The first element of this vector is the intercept, and the remaining parameters are called regression coefficients.

# Linear regression models

- In regression terminology,  $\epsilon = y - g(x)$  is called the **error**, which is a random variable assumed to be independent of  $x$ .
- We can therefore write the relationship between the response variable  $y$  and the explanatory variables  $x$  as follows:

$$y = g(x) + \epsilon$$

- For the observed data, we usually refer to the corresponding values of  $\epsilon$  as **residuals**.

## Least squares method

- There are many ways to estimate  $\beta$ , one of the most popular approach is the method of **least squares**, which is in general an optimization problem with no constraints

$$\text{minimize } ||y - x\beta||_2^2$$

- Recall that  $\ell_2$ -norm (Euclidean norm) is defined as

$$||z||_2 = (|z_1|^2 + |z_2|^2 + \dots + |z_n|^2)^{1/2}$$

In general, the  $\ell_p$  norm ( $p \geq 1$ ) is as follows:

$$||z||_p = (|z_1|^p + |z_2|^p + \dots + |z_n|^p)^{1/p}$$

- $||y - x\beta||_2^2 = \sum_{i=1}^n (y_i - x_i\beta)^2$  is called **residual sum of squares**, *RSS*, which is a quadratic function of regression parameters, *RRS*( $\beta$ ).

## Least squares method

- To find the value of  $\beta$  that minimizes  $RSS(\beta)$ , we set the first derivative to zero,

$$\frac{\partial RSS}{\partial \beta} = -2x'(y - x\beta)$$

$$\frac{\partial^2 RSS}{\partial \beta \partial \beta'} = 2x'x$$

- To have a unique solution for  $\beta$ ,  $x'x$  needs to be positive definite ( $x$  has to be full column rank).
- If this holds, the unique solution is obtained by setting the first derivative to zero

$$\begin{aligned} -2x'(y - x\beta) &= 0 \\ \hat{\beta} &= (x'x)^{-1}x'y \end{aligned}$$

## Sampling distribution of parameters

- So far, we have not made any assumption regarding the distributional form of the random variables (more specifically for the response variable since  $x$  is assumed to be fixed).
- We did not need to make such assumptions if all we wanted was point estimates of regression parameters.
- Usually, we want more than point estimates.
- We, for example, want to know about variability (e.g., standard error) of the estimates.



## Sampling distribution of parameters

- For this, we assume that  $x$  are fixed at the observed value and  $y$ 's are uncorrelated with a constant variance; i.e.,  $\text{cov}(y|x) = \sigma^2 I$  (note that we have not fully specified the distribution yet).
- As the result,

$$\begin{aligned}\text{cov}(\hat{\beta}) &= (x'x)^{-1}x'[(x'x)^{-1}x']'\sigma^2 \\ &= (x'x)^{-1}x'x(x'x)^{-1}\sigma^2 \\ &= (x'x)^{-1}\sigma^2\end{aligned}$$

- We also have

$$\begin{aligned}E(\epsilon) &= E(y) - E(E(y|x)) = E(y) - E(y) = 0 \\ \text{var}(\epsilon) &= \sigma^2\end{aligned}$$

## Estimating $\sigma$

- $\sigma$  itself is almost always unknown and needs to be estimated based on the data.
- To estimate  $\sigma$ , we usually use the following unbiased estimator

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y - x_i \hat{\beta})^2}{n - p - 1}$$

We use  $n - p - 1$  instead of  $n$  to make the estimate unbiased.

- The fit of the model can be measured based on  $\hat{\sigma}^2$ .
- For this, we use  $R^2 = 1 - \frac{\hat{\sigma}^2}{S_y^2}$ , which is the fraction of variance of response variable explained by the model. Here,  $S_y^2$  is the observed variance of  $y$ .

# Inference

- Note that while we could provide a measure of variability for the estimator of regression parameters, to perform statistical inference about these parameters, we need to make more assumptions about the distribution of  $y$ .
- We assume that

$$y|x, \beta, \sigma \sim N(x\beta, \sigma^2 I)$$

- Therefore,

$$\epsilon|\sigma \sim N(0, \sigma^2 I)$$

- As the result, we have

$$\begin{aligned}\hat{\beta}|\sigma &\sim N(\beta, (x'x)^{-1}\sigma^2) \\ \frac{n\hat{\sigma}^2}{\sigma^2} &\sim \chi^2(n-p-1)\end{aligned}$$

- Moreover, we can show that  $\hat{\beta}$  and  $\hat{\sigma}^2$  are independent.

# Inference

- Using the sampling distribution of  $\beta$ , we can obtain the confidence interval for a given confidence level  $c$ .
- For each individual  $\beta_j$  (corresponding to  $x_j$ ), the standard error is the square-root of the  $i^{th}$  element of the covariance matrix  $(x'x)^{-1}\sigma^2$ .
- The  $c$  level confidence interval for  $\beta_j$  can be obtained as

$$\hat{\beta}_j \pm t_c^* se(\hat{\beta}_j)$$

where  $t_c^*$  is the corresponding  $t$ -critical value based on  $t(n - p - 1)$  distribution.

## Inference

- To test the null hypothesis  $H_0 : \beta_j = 0$ , we can use the following  $T$ -statistics:

$$T = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

- Under  $H_0$ ,  $T$  has a  $\mathbf{t}(\mathbf{n} - \mathbf{p} - \mathbf{1})$  distribution.
- If we want to test the null hypothesis with respect to a group of coefficients, i.e.,  $H_0 : \beta_1 = \beta_2 = \dots, \beta_s = 0$  that reflects a reduced (simpler) model, we use the  $\mathbf{F}$  statistic

$$F = \frac{(RSS_r - RSS)/s}{RSS/(n - p - 1)} \sim \mathcal{F}(s, n - p - 1)$$

where  $RSS_r$  is the residual sum of squares for the reduced model.

## Likelihood function

- An alternative approach for estimating the parameters of linear regression model (and in general, most statistical models) is based on the **likelihood function**.
- To find the likelihood function, we first need to assume a probability distribution for the data, i.e.,  $P(y|\theta)$ , where  $\theta$  are unknown parameters.
- This distribution is based on our opinion regarding the mechanism that generates the data.
- The likelihood function is defined by plugging-in the observed data in the probability distribution and expressing it as a function of model parameters, i.e.,  $f(\theta, y)$ .

## Likelihood function

- For linear regression models, the data include the response variables  $y$  and the explanatory variables  $x$ . Therefore, in general we need to specify  $P(x, y)$ .
- However, since  $x$  are assumed to be fixed at their observed value,  $P(x) = 1$ , the joint distribution reduces to the conditional distribution of  $y|x$ .

$$P(x, y) = P(x)P(y|x) = P(y|x)$$

- Therefore, we only need to specify the conditional distribution of  $y$  given  $x$ .

## Likelihood function

- We assume this  $P(y|x)$  is a normal distribution.
- As we mentioned, we model the expectation of this distribution as a linear function of  $x$ , i.e.,  $E(y|x) = x\beta$ , and we assume the variance of this distribution is  $\sigma^2$  (which is independent of  $x$  and  $\beta$ ).
- Therefore, assuming that the observations are independent, we have

$$y|x, \beta \sim (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (y_i - x_i\beta)^2}{2\sigma^2}\right)$$

- The likelihood function is specified by plugging-in the observed values of  $x$  and  $y$  in the probability distribution and expressing the result as a function of  $\beta$  (for now, we assume  $\sigma$  is fixed).

$$f(\beta) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (y_i - x_i\beta)^2}{2\sigma^2}\right)$$



## Maximum likelihood estimation

- To estimate model parameters, we can find their values such that the probability of the observed data is maximum.
- For this, we maximize the likelihood function with respect to model parameters. Of course, it is easier to maximize the **log of likelihood function**, i.e.,  $L(\beta) = \log(f(\beta))$ .
- In general, this is a convex optimization problem.
- To maximize the likelihood function, we can focus on the part of the function that is related to the parameter (i.e., **kernel**).
- For linear regression models,

$$L(\beta) = -\sum_{i=1}^n (y_i - x_i\beta)^2 - \log(2\sigma^2)$$

## Maximum likelihood estimation

- For simplicity, we can also remove all the constant (not related to the parameters) parts;

$$L(\beta) = -\sum_{i=1}^n (y_i - x_i\beta)^2$$

- Now we can simply set the first derivative to zero (likelihood equation) to obtain the maximum likelihood estimate

$$\begin{aligned}\frac{\partial L(\beta)}{\partial \beta} &= 2\sum_{i=1}^n x_i(y_i - x_i\beta) \\ x'(y - x\beta) &= 0 \\ \hat{\beta} &= (x'x)^{-1}x'y\end{aligned}$$

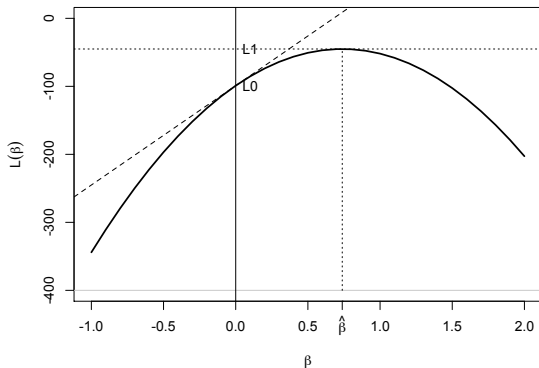
- In this case, MLE is the same as the least squares estimate.

# Maximum likelihood estimation

- Under weak regularity conditions, the MLE demonstrates attractive properties as  $n \rightarrow \infty$ : MLE is asymptotically
  - normal
  - consistent
  - efficient

# Maximum likelihood estimation

- This graphs shows the log-likelihood function and the location of MLE for randomly simulated data.



## Wald, score, and likelihood ratio tests

- Using the above graph, we can create three different tests for statistical inference:
  - **Wald**: based on the distance between  $\hat{\beta}$  and  $\beta_0$
  - **Score**: based on difference between the slopes at  $\hat{\beta}$  and  $\beta_0$
  - **Likelihood ratio**: based on the distance between  $L_1$  and  $L_0$

## Wald, score, and likelihood ratio tests

- Consider the null hypothesis  $H_0 : \beta = \beta_0$ , where  $\beta_0$  is the value of  $\beta$  under the null.
- Due to large-sample normality of MLE, we have

$$w = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$$

where  $w$  has an approximately  $N(0, 1)$  distribution.

- This type of statistics where we use the standard error of the estimator (as opposed to standard deviation of the null distribution) is referred to as **Wald statistic**.

## Wald, score, and likelihood ratio tests

- **Score test** on the other hand is based on the slope at  $\beta_0$ .
- This is in fact the value of **score function**,

$$u(\beta) = \partial L(\beta) / \partial \beta$$

evaluated at  $\beta_0$ .

- The dashed line in the above graph shows the slope at  $\beta_0 = 0$ .
- As we expect, the further  $\beta_0$  is away from the MLE, the larger this slope becomes in absolute value (i.e., we can reject the null hypothesis more confidently).

## Wald, score, and likelihood ratio tests

- The score statistic,  $s$ , is obtained by dividing the  $u(\beta_0)$  by its corresponding standard error.
- Under the null hypothesis:

$$s \sim N(0, 1)$$

- The advantage of score test is that we do not need to estimate the maximum likelihood estimate.



## Wald, score, and likelihood ratio tests

- The third test statistic is the likelihood ratio test.
- Here, we maximize the likelihood function under  $H_0$  and under  $H_0 \cup H_a$  (where  $H_a$  is the alternative hypothesis).
- The ratio of these two maximums is called the likelihood ratio test. In general,

$$LR = \frac{\sup_{\theta \in \Omega_0} f(\theta)}{\sup_{\theta \in \Omega} f(\theta)}$$

where  $\Omega_0$  is the parameter space under to  $H_0$ .

## Wald, score, and likelihood ratio tests

- In general, the likelihood ratio cannot exceed 1, since the maximized value under  $H_0$  would be less than or equal to the maximum value under  $H_0 \cup H_a$ .
- For hypothesis testing, we have  $-2 \log(LR) = -2(L_0 - L_1)$  has asymptotic  $\chi^2$  distribution with the degrees of freedom equal to the difference between the dimension of parameter space under  $H_0 \cup H_a$  and under  $H_0$ .
- Here  $L_1$  is the maximum value of log-likelihood under  $H_0 \cup H_a$ , and  $L_0$  is the maximum value of log-likelihood under  $H_0$ .
- For the simple linear regression, when testing the null hypothesis,  $H_0 : \beta = \beta_0$ ,  $L_1 = L(\hat{\beta})$  and  $L_0 = L(\beta_0)$ .
- $L_1$  and  $L_0$  (assuming  $H_0 : \beta = 0$ ) are shown in the above figure.

# What could go wrong with linear regression models

- In practice, one or more assumptions of linear regression models might be violated.
- This could result in wrong inference.
- Here, we discuss these assumptions.

## Linear relationship

- Using linear models, we implicitly assume that the relationship between  $x$  and  $y$  is **linear** (note that this is general different from the linearity assumption of the function in terms of parameters; i.e.,  $g(x) = x\beta$ ).
- If the assumption of linear relationship does not hold, we might still be able to use linear regression models after some transformation of original variables.
- Typical transformations are (we could use a combination of these with the original variables)
  - $\log(x)$ : For variables with positive values and heavily right-skewed distribution.
  - $\sqrt{x}$ : This transformation has milder effect compared to  $\log$  transformation, and it is usually recommended for counts.
  - $x^2, x^3, \dots$ : To create nonlinear relationships in the form of polynomial function.

## Additivity of effects

- In linear regression models, the effects of explanatory variables on the response variable are assumed to be additive.
- We could of course fix this by adding interactions terms
- We could also regress the log transformation of response variable on explanatory variables.
- This way, a model of the form,

$$\hat{y} = \beta_0 x_1^{\beta_1} x_2^{\beta_2}$$

where the effects are not additive, would become

$$\log(\hat{y}) = \log(\beta_0) + \beta_1 \log(x_1) + \beta_2 \log(x_2)$$

where the effects (on log scale) are additive

# Independence and constant variance assumptions for errors

- In linear regression models, the error terms are assumed to be independent.
- In this case, the covariance matrix of error terms is not diagonal ( $\sigma^2 I$ ) anymore, we need to use a full covariance matrix ( $\Sigma$ ).
- Moreover, they are assumed to have equal variance. When this is not the case, we could use **weighted least squares**, where the weight of each data point is inversely proportional to its variance.
- The assumption of normality for errors is not as important as the above two.

## Bounded response variable

- In linear regression analysis, we model the expected value of the response variable as a function of explanatory variables,  $E(y|x) = x\beta$ .
- The right hand side of this equation is unbounded in general. This could cause a problem, if the left hand side,  $E(y|x)$ , is bounded.
- For example, if the response variable is binary,  $y \in \{0, 1\}$ , its expectation is between 0 and 1.
- For count variables, the expectation would be a non-negative number.

# Generalized linear model

- To deal with some of these issues, we need a more flexible family of models.
- The class of generalized linear models (GLM), that includes linear models as a special case, provides such flexibility while it is still easy to use.
- Generalized linear models have three components:
  - A **random** component
  - A **systematic** component
  - A **link** function



## Generalized linear model

- The random component identifies the response variable and its probability distribution.
- In most situations, we assume parametric model  $P(y|\theta)$  for the distribution of  $y$  from the **exponential family**.
- Recall that the exponential family includes most of the well-known distributions such as normal, binomial, multinomial and Poisson.
- In general, if the outcome variable is continuous and real-valued, we use the **normal** distribution.
- If the outcome is binary, we use the **binomial** distribution. For outcome variables with multiple categories, we use the **multinomial** instead.
- If the outcome variable represent counts data, we use the **Poisson** distribution.

## Generalized linear model

- The systematic component specifies the set of predictors (i.e., explanatory variables)  $x = (x_1, \dots, x_p)$  used in a **linear predictor** function.
- As before, we also append a vector of ones at the beginning of  $x$ .
- In the matrix form, the linear predictor function  $\eta = x\beta$ , where  $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ .
- Alternatively, for each observation  $i$ , where  $i = 1, \dots, n$ , the linear predictor function is  $\eta_i = \beta_0 + \sum_j^p x_{ij}\beta_j$ .
- Also, as before, some of predictors could be a transformation (e.g.,  $x^2$ ) of original predictors.

## Generalized linear model

- The link function is a monotonic differentiable function that connects the random and systematic components.
- More specifically, if  $\mu = E(y|x)$ , the link function  $g^{-1}$  connects  $\mu$  to  $\eta$  such that  $g^{-1}(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j$  for each observation  $i$ .
- For the ordinary linear model we discussed before, the link function is identity:  $g^{-1}(\mu_i) = \mu_i$ . That is  $\mu_i = \eta_i = x_i\beta$ .

## Logistic regression model

- As mentioned before, for binary outcome variable, we use the binomial distribution.

$$y_i | n_i, \mu_i \sim \text{binomial}(n_i, \mu_i)$$

with the Bernoulli distribution as its special case when  $n_i = 1$ .

- As usual, we define the systematic part of the model  $\eta_i = x_i \beta$  (where  $x_i$  is a row vector of all observed values for subject  $i$ , and  $\beta$  is a column vector of size  $p + 1$ ).
- A common link function for this model is the **logit** function and defined as

$$g^{-1}(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = x_i \beta$$

where  $\mu_i$  is the probability of success (i.e.,  $y_i = 1$ ).

- As the result

$$\mu_i = \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)}$$

## Logistic regression model

- The likelihood is therefore defined in terms of  $\beta$  as follows:

$$p(y|\mu) \propto \prod_{i=1}^n \mu_i^{y_i} (1 - \mu_i)^{n_i - y_i}$$
$$p(y|\beta) \propto \prod_{i=1}^n \left( \frac{\exp(x_i \beta)}{1 + \exp(x_i \beta)} \right)^{y_i} \left( \frac{1}{1 + \exp(x_i \beta)} \right)^{n_i - y_i}$$

- Note that in this model the variance of  $y|x$  depends on the mean and therefore will not be constant

$$\text{var}(y_i|x_i) = \mu_i(1 - \mu_i)$$

## Interpretation

- To interpret  $\beta$ , note that

$$\log\left[\frac{P(y = 1|x, \beta)}{1 - P(y = 1|x, \beta)}\right] = (\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)$$

is the log of odds for the outcome of interest,  $y = 1$ .

- The intercept  $\beta_0$  is therefore the log of odds when the value of all covariates is 0.
- Alternatively,

$$\left[\frac{P(y = 1|x, \beta)}{1 - P(y = 1|x, \beta)}\right] = \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)$$

- Therefore, we can say,  $\exp(\beta_0)$  is the odds when all covariates are 0 (sometimes called the baseline odds).

## Interpretation

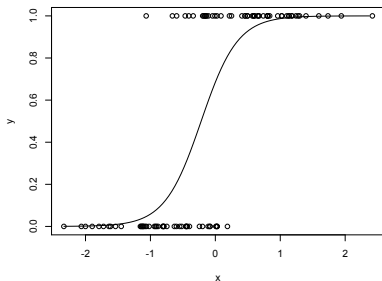
- $\exp(\beta_j)$  on the other hand is how much the odds multiplicatively increases for one unit increase in  $x_j$  when all other covariates are fixed.
- Or we can say,  $\exp(\beta_j)$  is the odds ratio for subjects with  $X_j = x_j + 1$  compared to subjects with  $X_j = x_j$  when all other covariates are fixed.
- Positive  $\beta_j$  indicates that the odds increases as  $x_j$  increases (everything else fixed), whereas for negative estimate of  $\beta_j$  the odds decreases as  $x_j$  increases (everything else fixed).

# Logistic regression model

- Positive  $\beta$

$$x_i \sim N(0, 1)$$

$$y_i \sim \text{Bernoulli}\left(\frac{\exp(1 + 3x)}{1 + \exp(1 + 3x)}\right)$$



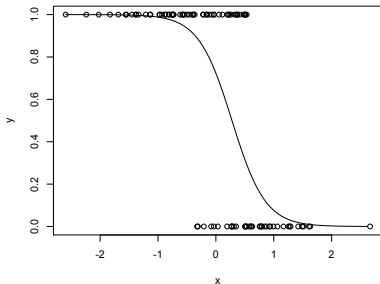


# Logistic regression model

- Negative  $\beta$

$$x_i \sim N(0, 1)$$

$$y_i \sim \text{Bernoulli}\left(\frac{\exp(1 - 3x)}{1 + \exp(1 - 3x)}\right)$$



## Logistic regression for retrospective studies

- Assume that we want to investigate the effect of smoking  $x_1$ , along with some other covariates  $(x_2, \dots, x_p)$  on lung cancer so we can build the following model

$$\log\left(\frac{\theta}{1-\theta}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (1)$$

where  $\theta = P(y = 1|x)$  is the probability of developing lung cancer.

- Obviously, we cannot assign subjects to “smoking” and “non-smoking” groups (clinical trials) to see who develops lung cancer, or let them decide their cohort and wait for a long time to see the results (cohort study).

## Logistic regression for retrospective studies

- Instead, we randomly select (with probability  $p_1$ ) some subjects from the population of patients with lung cancer, and randomly select some subjects (with probability  $p_0$ ) from the population of people without lung cancer. Then, we ask each subject whether he or she has been a smoker. That is, we design a **retrospective case-control** study.
- If we denote the decision to sample a subject as  $z = 1$  our sampling mechanism is according to the following rules:

$$P(z = 1|y = 0) = p_0$$

$$P(z = 1|y = 1) = p_1$$

- If we use this data in our logistic regression model, we are in fact modeling  $\phi = P(y = 1|z = 1, x)$ .

## Logistic regression for retrospective studies

- However, using Bayes' theorem and assuming that sampling procedure does not depend on  $x$ , we have

$$\begin{aligned}\phi &= P(y = 1|z = 1, x) = \\ &= \frac{P(z = 1|y = 1, x)P(y = 1|x)}{P(z = 1|y = 0, x)P(y = 0|x) + P(z = 1|y = 1, x)P(y = 1|x)} \\ &= \frac{p_1\theta}{p_0(\theta - 1) + p_1\theta}\end{aligned}$$

- Using Model (1), and plugging it  $\theta$  in the logistic regression model based on  $\phi$ , we have

$$\log\left(\frac{\phi}{1 - \phi}\right) = \beta_0^* + \beta_1x_1 + \dots + \beta_px_p$$

where  $\beta_0^* = \beta_0 + \log(p_1/p_0)$

- Therefore, the only difference between the two models would be the intercept; our inference about the effect of smoking on lung cancer,  $\beta_1$ , is valid even though it is based on a retrospective study.

## Multinomial logistic model

- This is a generalization of logistic regression when the outcome could have multiple values (i.e., could belong to one of  $K$  classes).

$$y_i | n_i, \mu_{i1}, \dots, \mu_{iK} \sim \text{multinomial}(n_i, \mu_{i1}, \dots, \mu_{iK})$$

where  $\mu_{ik}$  is the probability of class  $k$  for observation  $i$  such that  $\sum_{k=1}^K \mu_{ik} = 1$ .

- $y_i$  is also a vector of  $K$  elements with  $\sum_{k=1}^K y_{ik} = n_i$ .
- The systematic part is now a vector  $\eta_{ik} = x_i \beta$ , where  $\beta$  is a matrix of size  $(p+1) \times K$ .

## Multinomial logistic model

- Each column  $k$  (where  $k = 1, \dots, K$ ) corresponds to a set of  $p + 1$  parameters associated with class  $k$ .
- This representation is redundant and results in nonidentifiability, since one of the  $\beta_k$ 's (where  $k = 1, \dots, J$ ) can be set to zero without changing the set of relationships expressible with the model.
- Usually, either the parameters for  $k = 1$  (the first column) or for  $k = K$  (the last column) would be set to zero.

## Multinomial logistic model

- For the multinomial logistic model, we use a generalization of the link function we used for the binary logistic regression

$$\mu_{ik} = \frac{\exp(x_i \beta_k)}{\sum_{k'=1}^K \exp(x_i \beta_{k'})}$$

- The likelihood in terms of  $\beta$  is as follows:

$$p(y|\mu) \propto \prod_{i=1}^n \prod_{k=1}^K \mu_{ik}^{y_{ik}}$$
$$P(y|x, \beta) \propto \prod_{i=1}^n \prod_{k=1}^K \left( \frac{\exp(x_i \beta_k)}{\sum_{k'=1}^K \exp(x_i \beta_{k'})} \right)^{y_{ik}}$$

- Here  $\beta_k$  is a column vector of  $p + 1$  parameters corresponding to class  $k$ .

## Multinomial logistic model

- $\beta$  in general is a  $(p + 1) \times K$  matrix. The first row,  $(\beta_{01}, \dots, \beta_{0K})$  are intercepts, and  $(\beta_{j1}, \dots, \beta_{jK})$  in row  $j + 1$  are regression parameters associated with the  $j^{th}$  predictor.
- $x_i$  is the row vector of predictors value for observation  $i$  (including the constant 1 at the beginning).
- $y_{ik}$  is the number of cases in observation  $i$  that are in class  $k$ .



## Poisson model

- When the outcome variable,  $y$ , represents counts, we use the Poisson model.

$$y_i | \mu_i \sim \text{Poisson}(\mu_i)$$

- The systematic components are defined as before:  $\eta_i = x_i\beta$ .
- The usual link function for this model is the log link:

$$g^{-1}(\mu_i) = \log(\mu_i) = \eta_i$$

- We therefore have

$$\mu_i = \exp(\eta_i) = \exp(x_i\beta)$$

## Poisson model

- The likelihood in terms of  $\beta$  can be obtained as follows:

$$p(y_i|\mu_i) \propto \prod_i^n \exp(-\mu_i)\mu_i^{y_i}$$

$$p(y_i|\beta) \propto \prod_i^n \exp[-\exp(x_i\beta)][\exp(x_i\beta)]^{y_i}$$

- Similar to logistic and multinomial models, the variance of  $y|x$  in Poisson model depends on the mean and therefore will not be constant

$$\text{var}(y_i|x_i) = \mu_i$$

## Numerical methods for finding MLE

- As before, we maximize the likelihood function to estimate the regression parameters.
- We saw previously that the likelihood equation for linear regression models takes a simple form that can be solved analytically.
- However, the likelihood equations are in general nonlinear in  $\beta$ , and as the result, numerical methods are needed to find  $\hat{\beta}$ .
- A common approach is to use the Newton-Raphson method to maximize the likelihood function, or equivalently, maximize the log-likelihood function, or minimize the negative log-likelihood function (called the **energy function** in machine learning).

# Newton-Raphson method

- The Newton-Raphson method is a general purpose iterative algorithm for solving nonlinear equations.
- In statistics, we use this method to solve likelihood equation.
- Denote the log-likelihood as  $L(\beta)$ . Our objective is to find the value of  $\beta$  for which  $L(\beta)$  is maximized.
- We start with the single parameter case.

# Newton-Raphson method

- Start with an initial guess  $\beta^{(0)}$ .
- Iteratively update your guess as follows:
  - At each iteration  $n$ , use the Taylor series expansion (up to the quadratic term) around the current value of  $\beta^{(n)}$

$$L(\beta) \simeq L(\beta^{(n)}) + L'(\beta^{(n)})(\beta - \beta^{(n)}) + \frac{1}{2}L''(\beta^{(n)})(\beta - \beta^{(n)})^2$$

- Now take the derivative of  $L(\beta)$ , set it to zero and solve for  $\beta$ .
- Regard the answer as your next guess  $\beta^{(n+1)}$ :

$$\beta^{(n+1)} = \beta^{(n)} - \frac{L'(\beta^{(n)})}{L''(\beta^{(n)})}$$

- Continue the above process until the algorithm converges.

## Newton-Raphson method

- We can rewrite the equation for our next guess as

$$\beta^{(n+1)} = \beta^{(n)} + \frac{u(\beta^{(n)})}{o(\beta^{(n)})}$$

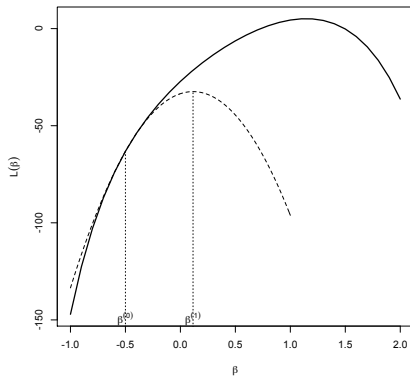
where  $u(\beta)$  is the score function, and

$$o(\beta^{(n)}) = -L''(\beta^{(n)})$$

- $o(\beta)$  is called the **observed information**, and its expectation  $i(\beta) = E[-L''(\beta)]$  is called the **Fisher information**.

# Newton-Raphson method

- The following graph illustrates how this method works.
- The solid line is the log-likelihood function,  $\beta^{(0)}$  is our initial guess, the dashed line is the approximate quadratic function around  $\beta^{(0)}$ , and  $\beta^{(1)}$  is our next guess.



## Multiple parameter

- For multiple parameter models (where  $\beta$  is a vector), we have

$$\beta^{(n+1)} = \beta^{(n)} + [o(\beta^{(n)})]^{-1} u(\beta^{(n)})$$

- where  $o(\beta)$  is a matrix whose  $(j, k)$  element is

$$o_{jk}(\beta) = -\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k}$$

- Therefore, the observed information matrix  $o(\beta)$  is the negative of the **Hessian matrix**.
- As before, the expected value of the  $o(\beta)$  is the Fisher information matrix.



# Fitting GLMs in R

- In practice, we use R and mainly the function `glm()` to fit generalized linear models.
- The function has the following format:

```
glm( formula, family = gaussian, data )
```

- `formula` specifies the systematic component, for example:

$$y \sim x_1 + x_2$$

$$y \sim .$$

## Fitting GLMs in R

- `family` This specifies the stochastic part of the model, i.e., probability distribution for the response variable.
- It could also specify the link function:

```
family = binomial(link = "logit").
```

- Some of the default links are

```
gaussian(link = "identity")  
binomial(link = "logit")  
poisson(link = "log")  
Gamma(link = "inverse")
```

- For multinomial logistic model, we can use the function `multinom` in the `nnet` package.
- If the categories are ordered, we can fit an ordinal logistic model using the function `polr()` in the `MASS` package.