Lecture 23

November 7, 2018

Maximum Likelihood in Logistic Regression

- We have binary responses $y_1, \ldots, y_n$ and data on p explanatory variables x_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Maximum Likelihood in Logistic Regression

- We have binary responses $y_1, \ldots, y_n$ and data on p explanatory variables x_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, p$.
- We assume that $y_1, \ldots, y_n$ are independent Bernoulli random variables with parameters $p_1, \ldots, p_n$.

Maximum Likelihood in Logistic Regression

- We have binary responses $y_1, \ldots, y_n$ and data on p explanatory variables $x_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, p$.
- We assume that $y_1, \ldots, y_n$ are independent Bernoulli random variables with parameters $p_1, \ldots, p_n$.
- We model the relationship between the response and explanatory variables by the formula

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \tag{1}$$

Maximum Likelihood in Logistic Regression

- We have binary responses $y_1, \ldots, y_n$ and data on p explanatory variables $x_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, p$.
- We assume that $y_1, \ldots, y_n$ are independent Bernoulli random variables with parameters $p_1, \ldots, p_n$.
- We model the relationship between the response and explanatory variables by the formula

$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \tag{1}$$

- Given data $y_1, \ldots, y_n$ and x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, how can be estimate the parameters $\beta_0, \ldots, \beta_p$.

- Note that the model can alternatively be written as

$$y_i \sim Ber\left(\frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}\right)$$

with $y_1, \ldots, y_n$ being independent.

- Note that the model can alternatively be written as

$$y_i \sim Ber\left(\frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}\right)$$

  with $y_1, \dots, y_n$ being independent.
- We use maximum likelihood to estimate $\beta_0, \dots, \beta_p$. The log-likelihood of the data $y_1, \dots, y_n$ (we take X to be deterministic) is

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{n} (y_i \log p_i + (1 - y_i) \log(1 - p_i)) \\ &= \sum_{i=1}^{n} \Big[y_i(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip}) - \\ &\log(1 + \exp\left(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip}\right)) \Big].\end{aligned}$$

- The MLE of β is the maximizer of $\ell(\beta)$.

- The MLE of β is the maximizer of $\ell(\beta)$.
- The maximizer of $\ell(\beta)$ cannot be computed in closed form. We use Newton's method for maximizing $\ell(\beta)$.

- The MLE of β is the maximizer of $\ell(\beta)$.
- The maximizer of $\ell(\beta)$ cannot be computed in closed form. We use Newton's method for maximizing $\ell(\beta)$.
- Newton's method uses the iterative scheme

$$\beta^{(m+1)} = \beta^{(m)} - \left(H\ell(\beta^{(m)})\right)^{-1} \nabla\ell(\beta^{(m)}) \tag{2}$$

where $\nabla\ell(\beta)$ and $H\ell(\beta)$ denote the gradient and Hessian of the function $\ell(\beta)$ respectively: $\nabla\ell(\beta) := (\partial\ell(\beta)/\partial\beta_0, \ldots, \partial\ell(\beta)/\partial\ell(\beta_p))^T$ and $H\ell(\beta)$ is the $(p+1) \times (p+1)$ matrix whose entries are second order derivatives of $\ell(\beta)$.

- For example, the $(1,1)$th entry of $H\ell(\beta)$ is $\partial^2\ell(\beta)/\partial\beta_0^2$, the $(1,2)$th entry is $\partial^2\ell(\beta)/\partial\beta_0\partial\beta_1$ and so on.

- For example, the $(1,1)$th entry of $H\ell(\beta)$ is $\partial^2\ell(\beta)/\partial\beta_0^2$, the $(1,2)$th entry is $\partial^2\ell(\beta)/\partial\beta_0\partial\beta_1$ and so on.
- We saw in the last class that

$$\nabla\ell(\beta) = X^T(Y-p) \quad \text{and} \quad H\ell(\beta) = -X^TWX$$

where W is the $n\times n$ diagonal matrix whose i^{th} diagonal entry is $p_i(1-p_i)$.

- For example, the $(1,1)$th entry of $H\ell(\beta)$ is $\partial^2\ell(\beta)/\partial\beta_0^2$, the $(1,2)$th entry is $\partial^2\ell(\beta)/\partial\beta_0\partial\beta_1$ and so on.
- We saw in the last class that

$$\nabla\ell(\beta) = X^T(Y-p) \quad \text{and} \quad H\ell(\beta) = -X^TWX$$

  where W is the $n \times n$ diagonal matrix whose i^{th} diagonal entry is $p_i(1-p_i)$.
- Newton's iterative scheme (2) therefore becomes

$$\beta^{(m+1)} = \beta^{(m)} + (X^TWX)^{-1}X^T(Y-p).$$

- For example, the $(1,1)$th entry of $H\ell(\beta)$ is $\partial^2\ell(\beta)/\partial\beta_0^2$, the $(1,2)$th entry is $\partial^2\ell(\beta)/\partial\beta_0\partial\beta_1$ and so on.
- We saw in the last class that

$$\nabla\ell(\beta) = X^T(Y - p) \quad \text{and} \quad H\ell(\beta) = -X^T W X$$

where W is the $n \times n$ diagonal matrix whose i^{th} diagonal entry is $p_i(1 - p_i)$.
- Newton's iterative scheme (2) therefore becomes

$$\beta^{(m+1)} = \beta^{(m)} + (X^T W X)^{-1} X^T (Y - p).$$

- This can be rewritten as

$$\beta^{(m+1)} = (X^T W X)^{-1} X^T W Z \tag{3}$$

- where

$$Z = X\beta^{(m)} + W^{-1}(Y - p). \tag{4}$$

- where
$$Z = X\beta^{(m)} + W^{-1}(Y - p). \tag{4}$$

- The method of estimating β therefore proceeds iteratively as follows.

- where
$$Z = X\beta^{(m)} + W^{-1}(Y - p). \tag{4}$$

- The method of estimating β therefore proceeds iteratively as follows.
- First have an initial estimate of $\beta_0, \ldots, \beta_p$.

- where

$$Z = X\beta^{(m)} + W^{-1}(Y - p). \tag{4}$$

- The method of estimating β therefore proceeds iteratively as follows.
- First have an initial estimate of $\beta_0, \ldots, \beta_p$.
- Call this initial estimator $\hat{\beta}^{(0)}$. Use this estimator to calculate p_i via

$$p_i = \frac{\exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} + \cdots + \hat{\beta}_p^{(0)} x_{ip})}{1 + \exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} \cdots + \hat{\beta}_p^{(0)} x_{ip})}.$$

- Use these values of p_i to create the response variable values Z_i via (6) and also use values of p_i to construct the matrix W. With Z and W, we can estimate β via

$$\hat{\beta}^{(1)} = (X^T W X)^{-1} X^T W Z.$$

- Use these values of p_i to create the response variable values Z_i via (6) and also use values of p_i to construct the matrix W. With Z and W, we can estimate β via

$$\hat{\beta}^{(1)} = (X^T WX)^{-1} X^T WZ.$$

- Now replace the initial estimator $\hat{\beta}^{(0)}$ by $\hat{\beta}^{(1)}$ and repeat this process. Keep repeating this until two successive estimates $\hat{\beta}^{(m)}$ and $\hat{\beta}^{(m+1)}$ do not change much. At that point, stop and report the estimate of β in the logistic regression model as $\hat{\beta}^{(m)}$.

- The expression $(X^T WX)^{-1} X^T WZ$ is reminiscent of the usual $(X^T X)^{-1} X^T Y$ which is the usual estimate of β in the linear model. In fact, this is the least squares estimate in a weighted least squares model as we shall describe next.

Weighted Least Squares

- Consider regression data in the usual set-up. Suppose we think that the right model is:

  $$Y = X\beta + e \quad \text{where,} \quad \mathbb{E}(e) = 0, \quad \text{and,} \quad Cov(e) = \sigma^2 V$$

  for some known (positive definite) matrix V.

Weighted Least Squares

- Consider regression data in the usual set-up. Suppose we think that the right model is:

$$Y = X\beta + e \quad \text{where,} \quad \mathbb{E}(e) = 0, \quad \text{and,} \quad Cov(e) = \sigma^2 V$$

  for some known (positive definite) matrix V.
- What then is a good estimator of β?

Weighted Least Squares

- Consider regression data in the usual set-up. Suppose we think that the right model is:

$$Y = X\beta + e \quad \text{where,} \quad \mathbb{E}(e) = 0, \quad \text{and,} \quad Cov(e) = \sigma^2 V$$

  for some known (positive definite) matrix V.
- What then is a good estimator of β?
- The difference from the usual situation is the presence of this matrix V.

Weighted Least Squares

- Consider regression data in the usual set-up. Suppose we think that the right model is:

$$Y = X\beta + e \quad \text{where,} \quad \mathbb{E}(e) = 0, \quad \text{and,} \quad Cov(e) = \sigma^2 V$$

  for some known (positive definite) matrix V.
- What then is a good estimator of β?
- The difference from the usual situation is the presence of this matrix V.
- It turns out the usual least squares estimator is not a good choice here for estimating β. It is better to use the weighted least squares estimator:

$$\hat{\beta}_{wls} := (X^T V^{-1} X)^{-1} X^T V^{-1} Y. \tag{5}$$

- It is not too hard to see that this estimator minimizes the weighted sum of squares

$$(Y - X\beta)^T V^{-1} (Y - X\beta)$$

over all β.

- It is not too hard to see that this estimator minimizes the weighted sum of squares

$$(Y - X\beta)^T V^{-1}(Y - X\beta)$$

  over all β.
- Why is it sensible to use (5) for estimating β in this case?

- It is not too hard to see that this estimator minimizes the weighted sum of squares

$$(Y - X\beta)^T V^{-1} (Y - X\beta)$$

over all β.
- Why is it sensible to use (5) for estimating β in this case?
- The follows reasons motivate this choice:

- It is not too hard to see that this estimator minimizes the weighted sum of squares

$$(Y - X\beta)^T V^{-1} (Y - X\beta)$$

  over all β.
- Why is it sensible to use (5) for estimating β in this case?
- The follows reasons motivate this choice:
- If e is multivariate normal, then (5) is the mle for β.

- Suppose V is diagonal. Then it is obvious that

$$(Y - X\beta)^T V^{-1} (Y - X\beta) = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2}{v_{ii}}$$

  where v_{ii} denotes the ith diagonal entry of V. It is intuitively clear that minimizing this weighted sum of squraes as opposed to the unweighted sum of squares is the right thing to do here.

- Suppose V is diagonal. Then it is obvious that

$$(Y - X\beta)^T V^{-1} (Y - X\beta) = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2}{v_{ii}}$$

  where v_{ii} denotes the ith diagonal entry of V. It is intuitively clear that minimizing this weighted sum of squraes as opposed to the unweighted sum of squares is the right thing to do here.
- For example, if v_{ii} is very high, it means that the ith observation is not very trustworthy and it therefore makes sense to give it low weight.

- Suppose V is diagonal. Then it is obvious that

$$(Y - X\beta)^T V^{-1} (Y - X\beta) = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2}{v_{ii}}$$

  where v_{ii} denotes the ith diagonal entry of V. It is intuitively clear that minimizing this weighted sum of squraes as opposed to the unweighted sum of squares is the right thing to do here.
- For example, if v_{ii} is very high, it means that the ith observation is not very trustworthy and it therefore makes sense to give it low weight.
- In the same way, it makes sense to give large weight to the ith observation if v_{ii} is low.

- One can show that $\hat{\beta}_{wls}$ is the BLUE for β.

- One can show that $\hat{\beta}_{wls}$ is the BLUE for β.
- The expectation and the covariance matrix of $\hat{\beta}_{wls}$ can be easily calculated via:

- One can show that $\hat{\beta}_{wls}$ is the BLUE for β.
- The expectation and the covariance matrix of $\hat{\beta}_{wls}$ can be easily calculated via:
- $$\mathbb{E}\hat{\beta}_{wls} = \beta \quad \text{and} \quad Cov(\hat{\beta}_{wls}) = \sigma^2(X^T V^{-1} X)^{-1}.$$

Iteratively Reweighed Least Squares for Logistic Regression Fitting

- Because of the similarity between (3) and (5), Newton's method for computing the maximum likelihood estimator in logistic regression can be seen as a sequence of weighted least squares estimators.

Iteratively Reweighed Least Squares for Logistic Regression Fitting

- Because of the similarity between (3) and (5), Newton's method for computing the maximum likelihood estimator in logistic regression can be seen as a sequence of weighted least squares estimators.
- That is why the iterative method is also called IRLS (Iteratively Reweighted Least Squares) or IWLS (Iteratively Weighted Least Squares).

Iteratively Reweighed Least Squares for Logistic Regression Fitting

- Because of the similarity between (3) and (5), Newton's method for computing the maximum likelihood estimator in logistic regression can be seen as a sequence of weighted least squares estimators.
- That is why the iterative method is also called IRLS (Iteratively Reweighted Least Squares) or IWLS (Iteratively Weighted Least Squares).
- Here is a more intuitive approach to understand IRLS. The goal is to fit the model (1) to the data.

Iteratively Reweighed Least Squares for Logistic Regression Fitting

- Because of the similarity between (3) and (5), Newton's method for computing the maximum likelihood estimator in logistic regression can be seen as a sequence of weighted least squares estimators.
- That is why the iterative method is also called IRLS (Iteratively Reweighted Least Squares) or IWLS (Iteratively Weighted Least Squares).
- Here is a more intuitive approach to understand IRLS. The goal is to fit the model (1) to the data.
- Because $p_i = \mathbb{E}y_i$, the equation (1) can be rewritten as

$$\log \frac{\mathbb{E}(y_i)}{1 - \mathbb{E}(y_i)} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- Because of the form above, a first idea to fit this model to data might be to try to fit a linear model to the response variable $\log(y_i/(1-y_i))$ on the explanatory variables and then to estimate $\beta_0, \ldots, \beta_p$ by the estimated coefficients of that linear model.

- Because of the form above, a first idea to fit this model to data might be to try to fit a linear model to the response variable $\log(y_i/(1-y_i))$ on the explanatory variables and then to estimate $\beta_0, \ldots, \beta_p$ by the estimated coefficients of that linear model.
- But because y_i is 0 or 1, the quantity $\log(y_i/(1-y_i))$ is either $-\infty$ or ∞ and so this response variable would make no sense.

- Because of the form above, a first idea to fit this model to data might be to try to fit a linear model to the response variable $\log(y_i/(1-y_i))$ on the explanatory variables and then to estimate $\beta_0, \dots, \beta_p$ by the estimated coefficients of that linear model.
- But because y_i is 0 or 1, the quantity $\log(y_i/(1-y_i))$ is either $-\infty$ or ∞ and so this response variable would make no sense.
- A way to fix this is to work with a response variable that is similar in spirit to $\log(y_i/(1-y_i))$ but which actually makes sense.

- Because of the form above, a first idea to fit this model to data might be to try to fit a linear model to the response variable $\log(y_i/(1-y_i))$ on the explanatory variables and then to estimate $\beta_0, \ldots, \beta_p$ by the estimated coefficients of that linear model.
- But because y_i is 0 or 1, the quantity $\log(y_i/(1-y_i))$ is either $-\infty$ or ∞ and so this response variable would make no sense.
- A way to fix this is to work with a response variable that is similar in spirit to $\log(y_i/(1-y_i))$ but which actually makes sense.
- Let $g(x) = \log(x/(1-x))$. By a first order Taylor expansion to g around p_i, we can write

$$g(y_i) \approx g(p_i) + g'(p_i)(y_i - p_i) = \log \frac{p_i}{1-p_i} + \frac{y_i - p_i}{p_i(1-p_i)}$$

- The right hand side above makes sense as opposed to $g(y_i)$. So we let

$$Z_i = \log \frac{p_i}{1 - p_i} + \frac{y_i - p_i}{p_i(1 - p_i)} \tag{6}$$

  and we can fit a linear model to Z_i based on the explanatory variables and estimate β by the estimated coefficients in that linear model.

- The right hand side above makes sense as opposed to $g(y_i)$. So we let

$$Z_i = \log \frac{p_i}{1-p_i} + \frac{y_i - p_i}{p_i(1-p_i)} \tag{6}$$

  and we can fit a linear model to Z_i based on the explanatory variables and estimate β by the estimated coefficients in that linear model.
- Should we estimate the coefficients of that linear model by ordinary least squares or should we use weighted least squares?

- The right hand side above makes sense as opposed to $g(y_i)$. So we let

$$Z_i = \log \frac{p_i}{1-p_i} + \frac{y_i - p_i}{p_i(1-p_i)} \tag{6}$$

  and we can fit a linear model to Z_i based on the explanatory variables and estimate β by the estimated coefficients in that linear model.
- Should we estimate the coefficients of that linear model by ordinary least squares or should we use weighted least squares? The variance of Z_i is:

$$var(Z_i) = var\left(\frac{y_i - p_i}{p_i(1-p_i)}\right) = \frac{1}{p_i(1-p_i)}.$$

- The right hand side above makes sense as opposed to $g(y_i)$. So we let

$$Z_i = \log \frac{p_i}{1-p_i} + \frac{y_i - p_i}{p_i(1-p_i)} \tag{6}$$

  and we can fit a linear model to Z_i based on the explanatory variables and estimate β by the estimated coefficients in that linear model.
- Should we estimate the coefficients of that linear model by ordinary least squares or should we use weighted least squares? The variance of Z_i is:

$$var(Z_i) = var\left(\frac{y_i - p_i}{p_i(1-p_i)}\right) = \frac{1}{p_i(1-p_i)}.$$

- Therefore if W is a diagonal matrix whose ith diagonal entry is $p_i(1-p_i)$, then

$$Cov(Z) = W^{-1}.$$

- Thus, while fitting a linear model to Z_i based on the explanatory variables, it is sensible to estimate the coefficients of the linear model by

$$(X^T W X)^{-1} X^T W Z.$$

- Thus, while fitting a linear model to Z_i based on the explanatory variables, it is sensible to estimate the coefficients of the linear model by
$$(X^T WX)^{-1} X^T WZ.$$
- This gives us the estimate of β in the logistic regression model:
$$\hat{\beta} := (X^T WX)^{-1} X^T WZ. \tag{7}$$

- Thus, while fitting a linear model to Z_i based on the explanatory variables, it is sensible to estimate the coefficients of the linear model by

$$(X^T W X)^{-1} X^T W Z.$$

- This gives us the estimate of β in the logistic regression model:

$$\hat{\beta} := (X^T W X)^{-1} X^T W Z. \tag{7}$$

- The obvious problem with the above approach is that we do not know p_i (p_i depends on the parameters $\beta_0, \dots, \beta_p$ that we are trying to estimate) and so we cannot really compute the response variable Z_i or the matrix W.

- The natural solution to this is to use an iterative method. First have an initial estimate of $\beta_0, \ldots, \beta_p$. Call this initial estimator $\hat{\beta}^{(0)}$. Use this estimator to calculate p_i via

$$p_i = \frac{\exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} + \cdots + \hat{\beta}_p^{(0)} x_{ip})}{1 + \exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} \cdots + \hat{\beta}_p^{(0)} x_{ip})}.$$

- The natural solution to this is to use an iterative method. First have an initial estimate of $\beta_0, \dots, \beta_p$. Call this initial estimator $\hat{\beta}^{(0)}$. Use this estimator to calculate p_i via

$$p_i = \frac{\exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} + \cdots + \hat{\beta}_p^{(0)} x_{ip})}{1 + \exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} \cdots + \hat{\beta}_p^{(0)} x_{ip})}.$$

- Use these values of p_i to create the response variable values Z_i via (6) and also use values of p_i to construct the matrix W. With Z and W, we can estimate β as in (7). Call this $\hat{\beta}^{(1)}$:

$$\hat{\beta}^{(1)} = (X^T W X)^{-1} X^T W Z.$$

- The natural solution to this is to use an iterative method. First have an initial estimate of $\beta_0, \ldots, \beta_p$. Call this initial estimator $\hat{\beta}^{(0)}$. Use this estimator to calculate p_i via

$$p_i = \frac{\exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} + \cdots + \hat{\beta}_p^{(0)} x_{ip})}{1 + \exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} \cdots + \hat{\beta}_p^{(0)} x_{ip})}.$$

- Use these values of p_i to create the response variable values Z_i via (6) and also use values of p_i to construct the matrix W. With Z and W, we can estimate β as in (7). Call this $\hat{\beta}^{(1)}$:

$$\hat{\beta}^{(1)} = (X^T W X)^{-1} X^T W Z.$$

- Now replace the initial estimator $\hat{\beta}^{(0)}$ by $\hat{\beta}^{(1)}$ and repeat this process. Keep repeating this until two successive estimates $\hat{\beta}^{(m)}$ and $\hat{\beta}^{(m+1)}$ do not change much. At that point, stop and report the estimate of β in the logistic regression model by $\hat{\beta}^{(m)}$.

- By what we have seen that this method is equivalent to computing the MLE by Newton's method.

Standard Errors for the MLE

- Is the MLE $\hat{\beta}$ unbiased (or at least approximately unbiased)?

Standard Errors for the MLE

- Is the MLE $\hat{\beta}$ unbiased (or at least approximately unbiased)? How do we compute its standard errors?

Standard Errors for the MLE

- Is the MLE $\hat{\beta}$ unbiased (or at least approximately unbiased)? How do we compute its standard errors? To answer these questions, consider the following simple heuristic argument.

Standard Errors for the MLE

- Is the MLE $\hat{\beta}$ unbiased (or at least approximately unbiased)? How do we compute its standard errors? To answer these questions, consider the following simple heuristic argument.
- Because $\hat{\beta}$ maximizes the loglikelihood, we have $\nabla \ell(\hat{\beta}) = 0$.

Standard Errors for the MLE

- Is the MLE $\hat{\beta}$ unbiased (or at least approximately unbiased)? How do we compute its standard errors? To answer these questions, consider the following simple heuristic argument.
- Because $\hat{\beta}$ maximizes the loglikelihood, we have $\nabla \ell(\hat{\beta}) = 0$. Let us now obtain a Taylor expansion of $\nabla \ell(\hat{\beta})$ around the true β:

$$0 = \nabla \ell(\hat{\beta}) \approx \nabla \ell(\beta) + H\ell(\beta) \left(\hat{\beta} - \beta \right).$$

Standard Errors for the MLE

- Is the MLE $\hat{\beta}$ unbiased (or at least approximately unbiased)? How do we compute its standard errors? To answer these questions, consider the following simple heuristic argument.
- Because $\hat{\beta}$ maximizes the loglikelihood, we have $\nabla \ell(\hat{\beta}) = 0$. Let us now obtain a Taylor expansion of $\nabla \ell(\hat{\beta})$ around the true β:

$$0 = \nabla \ell(\hat{\beta}) \approx \nabla \ell(\beta) + H\ell(\beta) \left(\hat{\beta} - \beta \right).$$

- Using the expressions $\nabla \ell(\beta) = X^T(Y - p)$ and $H\ell(\beta) = -X^T W X$, we obtain

$$0 \approx X^T(Y - p) - X^T W X(\hat{\beta} - \beta).$$

Standard Errors for the MLE

- Is the MLE $\hat{\beta}$ unbiased (or at least approximately unbiased)? How do we compute its standard errors? To answer these questions, consider the following simple heuristic argument.
- Because $\hat{\beta}$ maximizes the loglikelihood, we have $\nabla \ell(\hat{\beta}) = 0$. Let us now obtain a Taylor expansion of $\nabla \ell(\hat{\beta})$ around the true β:

$$0 = \nabla \ell(\hat{\beta}) \approx \nabla \ell(\beta) + H\ell(\beta) \left(\hat{\beta} - \beta \right).$$

- Using the expressions $\nabla \ell(\beta) = X^T(Y - p)$ and $H\ell(\beta) = -X^T W X$, we obtain

$$0 \approx X^T(Y - p) - X^T W X(\hat{\beta} - \beta).$$

- This gives

$$\hat{\beta} - \beta \approx (X^T W X)^{-1} X^T (Y - p).$$

- Because $\mathbb{E}Y = p$, this means that $\hat{\beta}$ is approximately unbiased for β.

- Because $\mathbb{E}Y = p$, this means that $\hat{\beta}$ is approximately unbiased for β. Also because $Cov(Y) = W$, we have

$$Cov(\hat{\beta}) \approx (X^T W X)^{-1}.$$

- Because $\mathbb{E}Y = p$, this means that $\hat{\beta}$ is approximately unbiased for β. Also because $Cov(Y) = W$, we have

$$Cov(\hat{\beta}) \approx (X^T WX)^{-1}.$$

- Therefore the approximate standard error of $\hat{\beta}_j$ is obtained by the square root of the corresponding diagonal entry of $(X^T WX)^{-1}$.