SHEET - An Introduction to Statistical Learning
Chapter 4 - Classification

Nicolas Playe

19 september 2024

1 Classification

1.1 Courses' Demonstrations

After a bit of manipulation of (4.2) we find that

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+X\beta_1} \tag{4.3}$$

Demonstration : $\frac{p(X)}{1-p(X)} = e^{\beta_0+X\beta_1}$

In the same way

We have

$$p(X) = \frac{e^{\beta_0+X\beta_1}}{1+e^{\beta_0+X\beta_1}} \tag{4.2}$$

$$\begin{aligned}\frac{p(X)}{1-p(X)} &= \frac{\frac{e^{\beta_0+X\beta_1}}{1+e^{\beta_0+X\beta_1}}}{1-\frac{e^{\beta_0+X\beta_1}}{1+e^{\beta_0+X\beta_1}}} \\ &= \frac{\frac{e^{\beta_0+X\beta_1}}{1+e^{\beta_0+X\beta_1}}}{\frac{1+e^{\beta_0+X\beta_1}-e^{\beta_0+X\beta_1}}{1+e^{\beta_0+X\beta_1}}} \\ &= \frac{\frac{e^{\beta_0+X\beta_1}}{1+e^{\beta_0+X\beta_1}}}{\frac{1}{1+e^{\beta_0+X\beta_1}}} \\ &= e^{\beta_0+X\beta_1}\end{aligned}$$

We arrive at

$$ln(\frac{p(X)}{1-p(X)}) = \beta_0 + X\beta_1 \tag{4.4}$$

Suppose that we wish to classify an observation into one of K classes, where $K \geq 2$. In other words, the qualitative response variable Y can take on K possible distinct and unordered values. Let π_k represent the overall or prior probability that a randomly chosen observation comes from the prior kth class. Let $f_k(X) = Pr(X|Y = k)$ 1 denote the density function of X density function for an observation that comes from the kth class. In other words, $f_k(x)$ is relatively large if there is a high probability that an observation in the kth class has $X = x$, and $f_k(x)$ is small if it is very unlikely that an observation in the kth class has $X = x$. Then Bayes' theorem states that

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{4.15}$$

Demonstration : $Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$

From Bayes theorem we have

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Then

$$\begin{aligned} Pr(Y = k|X = x) &= \frac{Pr(X = x|Y = k)Pr(Y = k)}{Pr(X = x)} \\ &= \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} Pr(Y = l)Pr(X = x|Y = l)} \\ Pr(Y = k|X = x) &= \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} \pi_l f_l(x)} \end{aligned}$$

The Bayes classifier involves assigning an observation X = x to the class for which (4.17) is largest. Taking the log of (4.17) and rearranging the terms, it is not hard to show that this is equivalent to assigning the observation to the class for which

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \tag{4.18}$$

is largest.

Demonstration : $\delta_k(x) = x.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$

We have

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)} \tag{4.17}$$

$$\begin{aligned} ln(p_k(x)) &= ln(\pi_k) - ln(\sqrt{2\pi\sigma}) - \frac{1}{2\sigma^2}(x-\mu_k)^2) \\ &- ln(\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{1}{2\sigma^2}(x-\mu_l)^2)) \end{aligned}$$

We can ignore the 2nde and last term for comparison, Then

$$\begin{aligned} ln(p_k(x)) &= ln(\pi_k) - \frac{1}{2\sigma^2}(x-\mu_k)^2) \\ &= ln(\pi_k) - -\frac{x^2}{2\sigma^2} - \frac{x\mu_k}{\sigma^2} + \frac{\mu_k^2}{\sigma^2} \end{aligned}$$

We can ignore the second term for comparaison, Then

$$\delta_k(x) = x.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

For instance, if K = 2 and $\pi_1 = \pi_2$, then the Bayes classifier assigns an observation to class 1 if $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$, and to class 2 otherwise. The Bayes decision boundary is the point for which $\delta_1(x) = \delta_2(x)$; one can show that this amounts to

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}. \tag{4.19}$$

Demonstration : $x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1+\mu_2}{2}$

Let

$$\delta_1(x) = \delta_2(x)$$

Then

$$\delta_1(x) - \delta_2(x) = 0$$

$$0 = x.\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log(\pi_1) - (x.\frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \log(\pi_2))$$

Let

$$\pi_1 = \pi_2$$

Then

$$\begin{aligned} 0 &= x.\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} - x.\frac{\mu_2}{\sigma^2} + \frac{\mu_2^2}{2\sigma^2} \\ &= x\frac{\mu_1 - \mu_2}{\sigma^2} + \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} \end{aligned}$$

$$\frac{\mu_1^2 - \mu_2^2}{2\sigma^2} = x\frac{\mu_1 - \mu_2}{\sigma^2}$$

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)}$$

$$x = \frac{(\mu_1 - \mu_2)(\mu_1 + \mu_2)}{2(\mu_1 - \mu_2)} \tag{4.19}$$

$$x = (\mu_1 + \mu_2) \tag{4.19}$$

Moreover we can show with the same reasonning that

$$\delta_1(x) \geq \delta_2(x)$$

implies

$$2x(\mu_1 - \mu_2) \geq \mu_1^2 - \mu_2^2$$

The LDA classifier plugs the estimates given in (4.20) and (4.21) into (4.18), and assigns an observation $X = x$ to the class for which

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k) \tag{4.22}$$

Demonstration : $\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$

By using $\hat{\sigma}$ instead of σ

By using $\hat{\mu}_k$ instead of μ_k

And by using $\hat{\pi_k}$ instead of π_k on the (4.18) formula

We get

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

In the case of p $>$ 1 predictors, the LDA classifier assumes that the observations in the kth class are drawn from a multivariate Gaussian distribution $N(\mu_k, \sum)$, where μ_k is a class-specific mean vector, and $\sum$ is a covariance matrix that is common to all K classes. Plugging the density function for the kth class, $f_k(X = x)$, into (4.15) and performing a little bit of algebra reveals that the Bayes classifier assigns an observation X $=$ x to the class for which

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \quad (4.2)$$

Demonstration : $\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k$

We have

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_k})\right) \tag{4.23}$$

Then

$$p_k(\mathbf{x}) = \frac{\frac{\pi_k}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_k}))}{\sum_{l=1}^{K}\frac{\pi_l}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_l})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_l}))}$$

$$\begin{aligned} ln(p_k(\mathbf{x})) &= ln(\pi_k) - ln((2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_k})) \\ &- ln(\sum_{l=1}^{K}\frac{\pi_l}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_l})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_l}))) \end{aligned}$$

We can ignore the 2nde and last term for comparison, Then

$$\begin{aligned} ln(p_k(\mathbf{x})) &= ln(\pi_k) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_k})) \\ &= -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k \end{aligned}$$

We can ignore the first term for comparison, Then

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k$$

As we have discussed, LDA assumes that the observations within each class are drawn from a multivariate Gaussian distribution with a class-specific mean vector and a covariance matrix that is common to all K classes. Quadratic discriminant analysis (QDA) provides an alternative approach. Like LDA, the QDA classifier results from assuming that the observations from each class are drawn from a Gaussian distribution, and plugging estimates for the parameters into Bayes' theorem in order to perform prediction. However, unlike LDA, QDA assumes that each class has its own covariance matrix. That is, it assumes that an observation from the kth class is of the form $X \sim N(\mu_k, \sum_k)$, where $\sum_k$ is a covariance matrix for the kth class. Under this assumption, the Bayes classifier assigns an observation X = x to the class for which

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log \pi_k$$

$$\delta_k(x) = -\frac{1}{2}x^T \Sigma_k^{-1} x + x^T \Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T \Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log \pi_k \quad (4.28)$$

Demonstration : $\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log \pi_k$

We have

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma_k}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top {\boldsymbol{\Sigma_k}}^{-1}(\mathbf{x}-\boldsymbol{\mu_k})\right) \quad (4.23)$$

Then

$$p_k(\mathbf{x}) = \frac{\frac{\pi_k}{(2\pi)^{p/2}|\boldsymbol{\Sigma_k}|^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top {\boldsymbol{\Sigma_k}}^{-1}(\mathbf{x}-\boldsymbol{\mu_k}))}{\sum_{l=1}^{K}\frac{\pi_l}{(2\pi)^{p/2}|\boldsymbol{\Sigma_l}|^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_l})^\top {\boldsymbol{\Sigma_l}}^{-1}(\mathbf{x}-\boldsymbol{\mu_l}))}$$

$$ln(p_k(\mathbf{x})) = ln(\pi_k) - ln((2\pi)^{p/2}) - \frac{1}{2}ln(|\boldsymbol{\Sigma_k}|)^{1/2}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top {\boldsymbol{\Sigma_k}}^{-1}(\mathbf{x}-\boldsymbol{\mu_k})$$
$$- ln(\sum_{l=1}^{K}\frac{\pi_l}{(2\pi)^{p/2}|\boldsymbol{\Sigma_l}|^{1/2}}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_l})^\top {\boldsymbol{\Sigma_l}}^{-1}(\mathbf{x}-\boldsymbol{\mu_l})))$$

We can ignore the 2nde and last term for comparison, Then

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log \pi_k$$

First, for LDA, we can make use of Bayes' theorem (4.15) as well as the assumption that the predictors within each class are drawn from amultivariate normal density (4.23) with class-specific mean and shared covariance matrix in order to show that

$$\log\left(\frac{\Pr(Y = k \mid X = x)}{\Pr(Y = K \mid X = x)}\right) = a_k + \sum_{j=1}^{p} b_{kj} x_j \tag{4.32}$$

Demonstration : $\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) = a_k + \sum_{j=1}^{p} b_{kj}x_j$

We have

$$\Pr(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}$$

Then

$$\begin{aligned}
\log(\frac{\Pr(Y=k \mid X=x)}{\Pr(Y=K \mid X=x)}) &= \log(\frac{\frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}}{\frac{\pi_K f_K(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}}) \\
&= \log(\frac{\pi_k f_k(x)}{\pi_K f_K(x)})
\end{aligned}$$

We have

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_k})\right) \tag{4.23}$$

Then

$$\begin{aligned}
\log(\frac{\Pr(Y=k \mid X=x)}{\Pr(Y=K \mid X=x)}) &= \log(\frac{\pi_k \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_k})\right)}{\pi_K \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_K})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu_K})\right)} \\
&= \log(\frac{\pi_k}{\pi_K}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top\mathbf{\Sigma}^{-1}(x-\mu_k) + \frac{1}{2}(x-\mu_K)^\top\Sigma^{-1}(x-\mu_K) \\
&= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mu_k+\mu_K)^T\Sigma^{-1}(\mu_k-\mu_K) + x^T\Sigma^{-1}(\mu_k-\mu_K) \\
&= a_k + x^T b_k \\
&= a_k + \sum_{j=1}^{p} b_{kj}x_j
\end{aligned}$$

Where

$$\begin{aligned}
a_k &= \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mu_k+\mu_K)^T\Sigma^{-1}(\mu_k-\mu_K) \\
b_{kj} &= [\Sigma^{-1}(\mu_k-\mu_K)]_j
\end{aligned}$$

Using similar calculations, in the QDA setting (4.31) becomes

$$\log\left(\frac{\Pr(Y=k \mid X=x)}{\Pr(Y=K \mid X=x)}\right) = a_k + \sum_{j=1}^{p} b_{kj}x_j + \sum_{j=1}^{p}\sum_{l=1}^{p} c_{kjl}x_jx_l, \tag{4.33}$$

Demonstration : $\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) = a_k+\sum_{j=1}^p b_{kj}x_j+\sum_{j=1}^p \sum_{l=1}^p c_{kjl}x_jx_l$

We have

$$\Pr(Y=k \mid X=x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

Then

$$\log(\frac{\Pr(Y=k \mid X=x)}{\Pr(Y=K \mid X=x)}) = \log(\frac{\frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}}{\frac{\pi_K f_K(x)}{\sum_{l=1}^K \pi_l f_l(x)}})$$

$$= \log(\frac{\pi_k f_k(x)}{\pi_K f_K(x)})$$

We have

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma_k}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top \boldsymbol{\Sigma_k}^{-1}(\mathbf{x}-\boldsymbol{\mu_k})\right) \tag{4.23}$$

Then

$$\log(\frac{\Pr(Y=k \mid X=x)}{\Pr(Y=K \mid X=x)}) = \log(\frac{\frac{\pi_k}{|\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top \boldsymbol{\Sigma_k}^{-1}(\mathbf{x}-\boldsymbol{\mu_k})\right)}{\frac{\pi_K}{|\Sigma_K|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_K})^\top \boldsymbol{\Sigma_K}^{-1}(\mathbf{x}-\boldsymbol{\mu_K})\right)}$$

$$= \log(\frac{\pi_k}{\pi_K}) - \frac{1}{2}log(|\Sigma_k|) + \frac{1}{2}log(|\Sigma_K|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top \boldsymbol{\Sigma_k}^{-1}(x-\mu_k)$$

$$+\frac{1}{2}(x-\mu_K)^\top \Sigma_K^{-1}(x-\mu_K)$$

$$= \log(\frac{\pi_k}{\pi_K}) - \frac{1}{2}log(|\Sigma_k|) + \frac{1}{2}log(|\Sigma_K|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top \boldsymbol{\Sigma_k}^{-1}(x-\mu_k)$$

$$+\frac{1}{2}(x-\mu_K)^\top \Sigma_K^{-1}(x-\mu_K)$$

$$= \log(\frac{\pi_k}{\pi_K}) - \frac{1}{2}log(|\Sigma_k|) + \frac{1}{2}log(|\Sigma_K|)$$

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu_k})^\top \boldsymbol{\Sigma_k}^{-1}(x-\mu_k)$$

$$+\frac{1}{2}(x-\mu_K)^\top \Sigma_K^{-1}(x-\mu_K)$$

By developping we have (be carreful of Σ_k and Σ_K dependences)

$$\log(\frac{\Pr(Y=k \mid X=x)}{\Pr(Y=K \mid X=x)}) = a_k + \sum_{j=1}^p b_{kj}x_j + \sum_{j=1}^p \sum_{l=1}^p c_{kjl}x_jx_l$$

Finally, we examine (4.31) in the naive Bayes setting. Recall that in this setting, f k (x) is modeled as a product of p one-dimensional functions $f_{kj}(x_j)$ for j=1,...,p. Hence,

$$\log\left(\frac{\Pr(Y=k\mid X=x)}{\Pr(Y=K\mid X=x)}\right) = a_k + \sum_{j=1}^{p} g_{kj}(x_j) \tag{4.34}$$

Demonstration : $\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right) = a_k + \sum_{j=1}^{p} g_{kj}(x_j)$

We have

$$\frac{\Pr(Y=k\mid X=x)}{\Pr(Y=K\mid X=x)} = \frac{\pi_k f_k(x)}{\pi_K f_K(x)}$$

$$f_k(\mathbf{x}) = \prod_{j=1}^{p} f_{kj}(x_j)$$

Then

$$\frac{\Pr(Y=k\mid X=x)}{\Pr(Y=K\mid X=x)} = \frac{\pi_k \prod_{j=1}^{p} f_{kj}(x_j)}{\pi_K \prod_{j=1}^{p} f_{Kj}(x_j)}$$

$$\log(\frac{\Pr(Y=k\mid X=x)}{\Pr(Y=K\mid X=x)} = \log\left(\frac{\pi_k}{\pi_K}\right) + \sum_{j=1}^{p}\log\left(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)}\right)$$

$$= a_k + \sum_{j=1}^{p} g_{kj}(x_j)$$

Where

$$a_k = \log\left(\frac{\pi_k}{\pi_K}\right)$$

$$g_{kj} = \log\left(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)}\right)$$

How does logistic regression tie into this story ? Recall from (4.12) that multinomial logistic regression takes the form

$$\log\left(\frac{\Pr(Y=k\mid X=x)}{\Pr(Y=K\mid X=x)}\right)=\beta_{k0}+\sum_{j=1}^{p}\beta_{kj}x_j$$

Demonstration : $\log\left(\frac{\Pr(Y=k|X=x)}{\Pr(Y=K|X=x)}\right)=\beta_{k0}+\sum_{j=1}^{p}\beta_{kj}x_j$

We have

$$\frac{\Pr(Y=k\mid X=x)}{\Pr(Y=K\mid X=x)}=\frac{\frac{e^{\beta_{k0}+\sum_{j=1}^{p}\beta_{kj}x_j}}{1+\sum_{l=1}^{K}e^{\beta_l 0+\sum_{j=1}^{p}\beta_{lj}x_j}}}{\frac{1}{1+\sum_{l=1}^{K}e^{\beta_l 0+\sum_{j=1}^{p}\beta_{lj}x_j}}}$$

$$\log\left(\frac{\Pr(Y=k\mid X=x)}{\Pr(Y=K\mid X=x)}\right)=\beta_{k0}+\sum_{j=1}^{p}\beta_{kj}x_j$$

Suppose that a random variable Y takes on nonnegative integer values, i.e. $Y \in 0, 1, 2, ...$ If Y follows the Poisson distribution, then

$$P(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \dots$$

Here, $\lambda \geq 0$ is the expected value of Y , i.e. E(Y). It turns out that λ also equals the variance of Y , i.e. $\lambda = E(Y) = Var(Y)$. This means that if Y follows the Poisson distribution, then the larger the mean of Y , the larger its variance.

Demonstration : $E(Y) = Var(Y)$

We have

$$P(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \dots$$

where $\lambda \geq 0$ is the waited value of$Y$.

$$\begin{aligned} E(Y) &= \sum_{k=0}^{\infty} kP(Y = k) \\ &= \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda}\lambda\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = e^{-\lambda}\lambda e^{\lambda} = \lambda. \end{aligned}$$

We have

$$\begin{aligned} Var(Y) &= E(Y^2) - (E(Y))^2 \\ E(Y^2) &= \sum_{k=0}^{\infty} k^2 P(Y = k) = \sum_{k=0}^{\infty}(k(k-1)P(Y = k)) + \sum_{k=0}^{\infty} kP(Y = k) \\ E(Y(Y-1)) &= \sum_{k=2}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\lambda^2\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} = \lambda^2 \\ E(Y^2) &= \lambda^2 + \lambda \\ Var(Y) &= E(Y^2) - (E(Y))^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda \\ Var(Y) &= E(Y) = \lambda \end{aligned}$$

To estimate the coefficients $\beta_0, \beta_1, ..., \beta_p$, we use the same maximum likelihood approach that we adopted for logistic regression in Section 4.3.2. Specifically, given n independent observations from the Poisson regression model, the likelihood takes the form

$$l(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!} \tag{4.38}$$

Demonstration : $l(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}$

We have

$$\mathcal{L}(\beta_0, \ldots, \beta_p) = \prod_{i=1}^{n} P(Y_i|X_i; \beta_0, \ldots, \beta_p)$$

$$P(Y = y_i) = \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}$$

$$\lambda(x_i) = e^{\beta_{i0}+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}}$$

Then

$$\mathcal{L}(\beta_0, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}$$

Equations (4.39)–(4.41) can be expressed using a link function, η, which applies a transformation to $E(Y|X_1, ..., X_p)$ so that the transformed mean is a linear function of the predictors. That is,

$$\eta(E(Y|X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{4.42}$$

The link functions for linear, logistic and Poisson regression are :
- for linear regression $\eta(\mu) = \mu$,
- for logistic regression $\eta(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
- for Poisson regression $\eta(\mu) = \log(\mu)$

Demonstration :
- for linear regression $\eta(\mu) = \mu$,
- for logistic regression $\eta(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
- for Poisson regression $\eta(\mu) = \log(\mu)$

We have

For linear regression We have

$E(Y|X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Finally

$\eta(\mu) = \mu,$

For logistic regression We have

$$ln(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{4.4}$$

Finally

$\eta(\mu) = log(\frac{\mu}{1 - \mu}),$

For Poisson regression We have

$E(Y|X_1, \ldots, X_p) = \lambda(X_1, \ldots, X_p) = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$

$log(E(Y|X_1, \ldots, X_p)) = log(\lambda(X_1, \ldots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Finally

$\eta(\mu) = log(\mu),$