Solution to Exercise 11.16

Consider the expression

$$-\log\left(\sum_{j=0}^{J} \exp(W_{tj}\beta^j)\right), \tag{11.89}$$

which appears in the loglikelihood function (11.35) of the multinomial logit model. Let the vector $\beta^j$ have $k_j$ components, let $k \equiv k_0 + \ldots + k_J$, and let $\beta \equiv [\beta^0 : \ldots : \beta^J]$. The $k \times k$ Hessian matrix $H$ of (11.89) with respect to $\beta$ can be partitioned into blocks of dimension $k_i \times k_j$, $i = 0, \ldots, J$, $j = 0, \ldots, J$, containing the second-order partial derivatives of (11.89) with respect to an element of $\beta^i$ and an element of $\beta^j$. Show that, for $i \neq j$, the $(i, j)$ block can be written as

$$p_i p_j W_{ti}^T W_{tj},$$

where $p_i \equiv \exp(W_{ti}\beta^i)/(\sum_{j=0}^{J} \exp(W_{tj}\beta^j))$ is the probability ascribed to choice $i$ by the multinomial logit model. Then show that the diagonal $(i, i)$ block can be written as

$$-p_i(1 - p_i)W_{ti}^T W_{ti}.$$

Let the $k$-vector $a$ be partitioned conformably with the above partitioning of the Hessian $H$, so that we can write $a = [a_0 : \ldots : a_J]$, where each of the vectors $a_j$ has $k_j$ components for $j = 0, \ldots, J$. Show that the quadratic form $a^T Ha$ is equal to

$$\left(\sum_{j=0}^{J} p_j w_j\right)^2 - \sum_{j=0}^{J} p_j w_j^2, \tag{11.90}$$

where the scalar product $w_j$ is defined as $W_{tj}a_j$.

Show that expression (11.90) is nonpositive, and explain why this result shows that the multinomial logit loglikelihood function (11.35) is globally concave.

The derivative of (11.89) with respect to the vector $\beta^i$ is the column vector

$$-\exp(W_{ti}\beta^i)/(\sum_{j=0}^{J} \exp(W_{tj}\beta^j)) W_{ti}^T. \tag{S11.12}$$

The derivative of the numerator with respect to $\beta^j$, for $j \neq i$, is 0. Therefore, using the definition of $p_i$, the derivative of (S11.12) with respect to $\beta^j$ is

$$\frac{\exp(W_{ti}\beta^i) \exp(W_{ti}\beta^j)}{(\sum_{i=0}^{J} \exp(W_{ti}\beta^i))(\sum_{j=0}^{J} \exp(W_{tj}\beta^j))} W_{ti}^T W_{tj} = p_i p_j W_{ti}^T W_{tj}.$$
Similarly, the derivative of (S11.12) with respect to \( \beta^i \) is

\[
- \exp(W_{ti} \beta^i) \sum_{j=0}^{J} \exp(W_{ti} \beta^i) + \exp^2(W_{ti} \beta^i) \frac{W_{ti}^\top W_{ti}}{\left( \sum_{j=0}^{J} \exp(W_{ti} \beta^i) \right)^2}
\]

\[= -p_i(1 - p_i)W_{ti}^\top W_{ti}.
\]

This completes the first part of the exercise.

The quadratic form in which we are interested is

\[a^\top Ha = J \sum_{j=0}^{J} \sum_{i \neq j} a^\top_i H_{ij} a_j,
\]

where we obtained formulas for the diagonal and off-diagonal blocks of \( H \) in the first part of the exercise. Using these, for \( i \neq j \) we have

\[a^\top_i H_{ij} a_j = p_i p_j a^\top_i W_{ti}^\top W_{tj} a_j = p_i p_j w_i w_j,
\]

(S11.13)

and for \( i = j \) we have

\[a^\top_i H_{ii} a_i = -p_i(1 - p_i)W_{ti}^\top W_{ti} a_i = -p_i(1 - p_i)w_i^2.
\]

(S11.14)

Thus the entire quadratic form is

\[
\sum_{j=0}^{J} \sum_{i \neq j} p_i p_j w_i w_j + \sum_{j=0}^{J} (p_j^2 w_j^2 - p_j w_j^2).
\]

This is equal to expression (11.90), because

\[
\left( \sum_{j=0}^{J} p_j w_j \right)^2 - \sum_{j=0}^{J} p_j w_j^2
\]

\[= \sum_{j=0}^{J} p_j^2 w_j^2 + \sum_{j=0}^{J} \sum_{i \neq j} p_i p_j w_i w_j - \sum_{j=0}^{J} p_j w_j^2
\]

\[= \sum_{j=0}^{J} \sum_{i \neq j} p_i p_j w_i w_j + \sum_{j=0}^{J} (p_j^2 w_j^2 - p_j w_j^2).
\]

This completes the second part of the exercise.

In order to show that expression (11.90) is nonpositive, we can use the Cauchy-Schwartz inequality, expression (2.08). Let the \((J + 1)\)-vectors \( x \) and \( y \) have typical elements \( p_{j}^{1/2} \) and \( p_{j}^{1/2} w_j \), respectively. Then

\[
\|x\|^2 = \sum_{j=0}^{J} p_j = 1, \quad \|y\|^2 = \sum_{j=0}^{J} p_j w_j^2, \quad \text{and} \quad x^\top y = \sum_{j=0}^{J} p_j w_j.
\]
Thus expression (11.90) can be rewritten as

\[ |x^\top y|^2 - \|x\|^2 \|y\|^2. \]

The Cauchy-Schwarz inequality tells us that \( |x^\top y|^2 \leq \|x\|^2 \|y\|^2 \), and this immediately tells us that expression (11.90) must be nonpositive.

This result shows that the multinomial logit loglikelihood function (11.35) is globally concave, because \( H \) is the Hessian of the loglikelihood function as well as the Hessian of expression (11.89). The first term in (11.35) is linear in the parameter vectors \( \beta^j \), and so it does not contribute to the Hessian at all.