USE OF BOX-COX TRANSFORMATION WITH BINARY RESPONSE MODELS

by

Victor M. Guerrero  Richard A. Johnson
University of Wisconsin

DISTRIBUTION STATEMENT A
Approved for public release; Distribution Unlimited
Use of the Box-Cox Transformation with Binary Response Models

by

Victor M. Guerrero
University of Wisconsin

Richard A. Johnson
University of Wisconsin

Summary

The power transformation suggested by Box and Cox (1964) is applied to the odds ratio to generalize the logistic model and to parameterize a certain type of lack of fit. Transformation of the design variable within the context of the dose-response problem is also considered.

1. Introduction

The use of linear logistic regression models is now widespread (c.f. Cox (1970), Merlove and Press (1973)). By introducing an additional parameter that allows for other than logarithmic transformations of the odds-ratio, we extend their applicability. In the spirit of Box and Cox (1964), our transformations are data based. Viewed in another way, by determining plausible values for the transformation parameter, we are able to decide whether or not the logarithm is an appropriate transformation. In this sense, we obtain a single parameter lack of fit criterion for linear logistic models.

As suggested by Cox (1970), we also consider transformations of the independent variable in the context of the dose-response problem. In this case, we obtain the correct asymptotic covariance matrix of the estimators by allowing the assumed model to be incorrect.

2. Linear Models for Proportions

In the linear logistic regression model, the assumption is made that a linear relationship is appropriate for linking the log-odds ratio of the dependent variable to several explanatory variables.

That is, if \( Y_1, \ldots, Y_n \) are independent 0-1 random variables with

\[ P_j = P(Y_j = 1), \]

then

\[ \log \left( \frac{P_j}{1 - P_j} \right) = \beta^T X_j, \quad j = 1, \ldots, n \]

where \( \beta \) is a q-dimensional vector of unknown parameters and

\( X_j = (1, X_{j1}, \ldots, X_{j(q-1)})' \)

is the \( j \)th vector of observations on \((q-1)\) explanatory variables.

Alternative transformations of the probability \( p_j \) have been suggested for linearizing purposes. Among the most common alternatives are the integrated normal, the arc-sine and the identity transformations. However, the arc-sine and the identity have finite ranges which sometimes limits their usefulness. On the other hand, the choice between the logistic and the normal functions is usually a matter of taste although, in recent times, the logistic model seems to have more advocates (c.f. Merlove and Press (1973)). Prentice (1976a) gives some reasons why the odds-ratio should be particularly considered in retrospective studies. Also, as pointed out by Cox (1970, p. 26), differences on a logistic scale have simpler interpretation in terms of the odds for success against failure.

The previous considerations have led us to study a "natural" extension of the linear logistic regression model. We assume that some power transformation of the odds-ratio satisfies a linear model. That is,

\[ \left( \frac{p_j}{1 - p_j} \right)^{(1)} = \beta^T X_j \quad \text{for} \quad j = 1, \ldots, n \]
where \( \theta \) and \( \pi_j \) are as in (1) and
\[
\left( \frac{p_j}{1-p_j} \right) = \begin{cases} 
\log(1-p_j) & \text{if } \lambda = 0 \\
\frac{1}{\lambda} \left( \frac{p_j}{1-p_j} \right)^{\lambda} - 1 & \text{if } \lambda \neq 0.
\end{cases}
\]

Model (3) focuses again on the odds ratio, includes as a special case the logistic model and can be used in all situations in which logistic regression is generally employed. One thing that should be remembered is that one extra degree of freedom will be used in estimating the transformation parameter \( \lambda \).

In the remainder of this section, we will assume that there exist \( k \) different conditions at which successes are recorded. Let \( n_k \) be the number of observations at condition \( k \) and \( r_k \) the number of successes (\( i = 1, \ldots, k \)). We tentatively assume that there is some value for \( \lambda_0 \) such that (2) holds. Then
\[
P_k(\theta_0) = \begin{cases} 
\left( 1 + \exp(-\theta_0 X_k) \right)^{-1} & \text{if } \lambda_0 = 0 \\
1 + (1 + \lambda_0 X_k)^{1-\lambda_0} & \text{if } \lambda_0 \neq 0
\end{cases}
\]

for some parameter values \( \theta_0 = (\theta_0^1, \ldots, \theta_0^k) \).

The likelihood function for \( n = \sum_{i=1}^{k} p_k \) observations is given by
\[
L_k(\theta) = \prod_{i=1}^{k} \left( \frac{n_k}{p_k} \right)^{r_k} p_k^{r_k} (1 - p_k)^{n_k - r_k}
\]

which yields the log-likelihood
\[
L_k(\theta) = c + \sum_{i=1}^{k} \left( r_k \log(p_k(\theta)) + (n_k - r_k) \log(1 - p_k(\theta)) \right)
\]

or
\[
L_k(\theta) = \begin{cases} 
c + \sum_{i=1}^{k} r_k \theta_i - \sum_{i=1}^{k} n_k \log(1 + \exp(\theta_i X_k)) & \text{if } \lambda = 0 \\
c + \sum_{i=1}^{k} r_k \theta_i \log(1 + \lambda \theta_i X_k) - \sum_{i=1}^{k} n_k \log(1 + (1 + \lambda \theta_i X_k)^{1/\lambda}) & \text{if } \lambda \neq 0
\end{cases}
\]

with \( c = \sum_{i=1}^{k} \log(n_k) - \log(r_k) - \log(n_k - r_k) \).

Even when model (3) is not correct, we are able to establish the strong consistency of the MLE.

**Theorem 1:** Let \( p_k(\theta) \) be given by (3) and \( p_k \) be the unknown true probability of success. Suppose that

1. the parameter space is a compact subset of \( \mathbb{R}^{k+1} \)
2. \( \lim_{n \to \infty} s_k = s_k \) with \( s_k \in (0, 1) \) and \( \sum_{k=1}^{k} s_k = 1 
3. \( H(\theta) = k \sum_{k=1}^{k} s_k \log \left( \frac{p_k(\theta)}{p_k} \right) + (1 - s_k) \log \left( \frac{1 - p_k(\theta)}{1 - p_k} \right) \) has a unique global maximum at \( \theta = \theta_0 \).

Then, \( \lim_{n \to \infty} \hat{\theta} = \theta_0 \) with probability one.
Proof: Let us write \( \hat{a}_k = \frac{a_k}{n_k} = \frac{1}{n_k} \sum_{j=1}^{n_k} Y_{ij} a_k \). Then, almost surely
\[
P_k \rightarrow P_k \text{ as } a_k \rightarrow 0 \text{, for } k = 1, \ldots, k.
\]
Applying Stirling's formula for factorials, we have
\[
\log(n_1) - \log(r_1) - \log(|r_1 - r_2| - 1) = -n_k \hat{a}_k \log(\hat{p}_k) + \log(1 - \hat{p}_k) + o(n_k^{-1})
\]
on an almost sure set, where \( O(n_k^{-1}) \) is uniform in \( \theta \). So, it follows that
\[
\frac{1}{n} \ln n(\hat{\theta}) = \sum_{k=1}^{k} \left[ (\log \hat{p}_k (\hat{\theta})) + (1 - \hat{p}_k) \log(1 - \hat{p}_k) + \log(1 - \hat{p}_k) + o(1) \right]
\]
with probability one.

Now, from (4), \( l(n) \) is seen to be continuous in \( \theta \) and \( \lambda \).

By compactness of the parameter space and continuity of \( l(n) \), we obtain, with probability one
\[
\lim_{n \to \infty} \mathbb{E}[\ln \hat{p}_k (\hat{\theta})] = \sum_{k=1}^{k} \left[ \frac{1}{n_k} \ln \hat{p}_k (\hat{\theta}) + (1 - \hat{p}_k) \ln \frac{1}{1 - \hat{p}_k} \right]
\]
uniformly in \( \theta \). Because the limit has a maximum at \( \theta_0 > 0 \), \( \hat{\theta}_n = \hat{\theta}_0 \) as \( n \to \infty \).

Remark: For each \( k \), the Kullback-Leibler information number between the true probability distribution and model (3) is
\[
K_k \left( \log \frac{P_k (\theta)}{P_k (\hat{\theta})} \right) = P_k \left( \log \frac{P_k (\theta)}{P_k (\hat{\theta})} + (1 - P_k) \log \frac{1 - P_k}{1 - P_k} \right)
\]
Then, we notice that
\[
\lim_{n \to \infty} \mathbb{E}[\ln \hat{p}_k (\hat{\theta})] = \sum_{k=1}^{k} \left[ \frac{1}{n_k} \ln \hat{p}_k (\hat{\theta}) + (1 - \hat{p}_k) \ln \frac{1}{1 - \hat{p}_k} \right],
\]
Thus, maximizing the log-likelihood under model (3) is asymptotically equivalent to minimizing the Kullback-Leibler information number between the true and the proposed models.

The asymptotic normality of \( \hat{\theta}_n \) is stated in the following theorem.

Theorem 2: If (1) holds for some \( \theta_1 = (\theta_1, \lambda) \)
\[
\sqrt{n} (\hat{\theta}_n - \theta_0) \rightarrow N_{q+1} \left( 0, \Sigma \right) \text{ as } n \to \infty \text{, for } \theta_1 < 1 \text{.}
\]
Where
\[
v^{-1} = \exp \left( \sum_{k=1}^{k} \left[ \frac{1}{n_k} \ln \hat{p}_k (\hat{\theta}) + (1 - \hat{p}_k) \ln \frac{1 - \hat{p}_k}{1 - \hat{p}_k} \right] \right)
\]
We suggest to obtain the MLE’s by solving the likelihood equations for \( \hat{\theta}_k \) at a fixed value of \( \lambda \) and then varying \( \lambda \) until the log-likelihood is maximized. The normal equations for \( \theta_1, \ldots, \theta_q \) are
\[
\sum_{k=1}^{k} \left[ X_{ik} (1 + \lambda) \hat{\theta}_k \right] X_{ik}^\prime (r_{i} - n_k \hat{p}_k (\hat{\theta})) = 0, \text{ for } k = 1, \ldots, q.
\]
A consistent estimator of the variance-covariance matrix of \( \hat{\theta}_n \) can be
obtained by inverting \( \frac{\partial^2 \ln L}{\partial \beta_i \partial \beta_j} \)
and evaluating it at

\[ \hat{\beta}_i = \hat{\beta}_i \]

where

\[ \frac{\partial^2 \ln L}{\partial \beta_i \partial \beta_j} = \sum_{r=1}^{k} X_{rj}(1 + \lambda \beta_i X_{rj})^{-2} \left( \lambda \beta_j X_{rj} (1 + \lambda \beta_i X_{rj}) \right) p_1(\theta) \]

\[ - \lambda^{-1} \frac{\partial^2 (\lambda g^i X_j)}{\partial \beta_i \partial \beta_j} + \lambda \left( \frac{1}{\lambda g^i X_j} \right)^{2} \left( \lambda \beta_j X_{rj} (1 + \lambda \beta_i X_{rj}) \right) \]

\[ + \lambda \left( \frac{1}{\lambda g^i X_j} \right)^{2} \left( \lambda \beta_j X_{rj} (1 + \lambda \beta_i X_{rj}) \right) \]

\[ \lambda \neq 0, \quad u, v = 1, \ldots, k \]

3. A 2x2 factorial arrangement

The situation we are considering covers the 2x2^3 factorial system which arises frequently in either designed experiments or survey studies, when the measurement of interest is the proportion of units with certain characteristics. The example that follows may be considered "classical" in the sense that many authors writing on categorical data have studied it (c.f. Cox (1970), Flieberg (1977)).

Dyke and Patterson (1952) were the first to perform a maximum likelihood analysis on data collected by Lombard and Doering. Four factors were considered important in affecting the probability of getting a good score in a test on cancer knowledge: The factors being:

(a) newspapers, (b) radio, (c) solid reading, and (d) lectures.

In the study, the aim was to estimate the main effects. Thus, we considered a model containing only six parameters, namely \( \lambda, \beta_1 \) (overall mean), \( \beta_2 \) (newspapers), \( \beta_3 \) (radio), \( \beta_4 \) (solid reading) and \( \beta_5 \) (lectures). The MLE of \( \theta \) was obtained by maximizing the log-likelihood first with respect to \( \beta_i \) for fixed \( \lambda \), and then searching for a maximum over the values of \( \lambda \). Figure 1 shows the graph of maximized log-likelihood function. The value of \( \lambda \), as read off the graph is .425, with a 95% confidence interval from -.112 to 1.104. Even though the value of \( \lambda \) is not significantly different from zero, we perform the analysis on this new scale. The original data and results of the analysis are presented in Table 1.

It should be noticed that \( \lambda \) is selected in accordance to the model proposed. Thus \( \hat{\lambda} \) obtained under a model which takes into account only main effects, is not necessarily the best scale for a model which also contains interactions. To illustrate this point we introduce the first order interactions of lectures with the three
## Table 1

**CLASSIFICATION OF INDIVIDUALS WITH RESPECT TO CANCER KNOWLEDGE**
*(Taken from Dyke and Patterson (1952))*

<table>
<thead>
<tr>
<th>Factor Combination</th>
<th>No. of Trials</th>
<th>No. with Good Scores</th>
<th>Observed Proportion</th>
<th>Estimated Proportion ($\hat{\lambda} = .425$)</th>
<th>Contribution to Chi-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>477</td>
<td>84</td>
<td>.176</td>
<td>.174</td>
<td>.010</td>
</tr>
<tr>
<td>a</td>
<td>231</td>
<td>75</td>
<td>.325</td>
<td>.321</td>
<td>.007</td>
</tr>
<tr>
<td>b</td>
<td>63</td>
<td>13</td>
<td>.206</td>
<td>.244</td>
<td>.277</td>
</tr>
<tr>
<td>c</td>
<td>150</td>
<td>67</td>
<td>.447</td>
<td>.413</td>
<td>.266</td>
</tr>
<tr>
<td>d</td>
<td>12</td>
<td>2</td>
<td>.167</td>
<td>.292</td>
<td>.459</td>
</tr>
<tr>
<td>ab</td>
<td>94</td>
<td>35</td>
<td>.372</td>
<td>.392</td>
<td>.058</td>
</tr>
<tr>
<td>ac</td>
<td>378</td>
<td>201</td>
<td>.532</td>
<td>.543</td>
<td>.043</td>
</tr>
<tr>
<td>ad</td>
<td>13</td>
<td>7</td>
<td>.538</td>
<td>.437</td>
<td>.171</td>
</tr>
<tr>
<td>bc</td>
<td>32</td>
<td>16</td>
<td>.550</td>
<td>.481</td>
<td>.013</td>
</tr>
<tr>
<td>bd</td>
<td>7</td>
<td>4</td>
<td>.571</td>
<td>.364</td>
<td>.523</td>
</tr>
<tr>
<td>cd</td>
<td>11</td>
<td>3</td>
<td>.273</td>
<td>.521</td>
<td>.623</td>
</tr>
<tr>
<td>abc</td>
<td>169</td>
<td>102</td>
<td>.604</td>
<td>.596</td>
<td>.007</td>
</tr>
<tr>
<td>abd</td>
<td>12</td>
<td>8</td>
<td>.667</td>
<td>.501</td>
<td>.330</td>
</tr>
<tr>
<td>acd</td>
<td>45</td>
<td>27</td>
<td>.600</td>
<td>.627</td>
<td>.020</td>
</tr>
<tr>
<td>bcd</td>
<td>4</td>
<td>1</td>
<td>.250</td>
<td>.576</td>
<td>.313</td>
</tr>
<tr>
<td>abcd</td>
<td>31</td>
<td>23</td>
<td>.742</td>
<td>.670</td>
<td>.080</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1729</strong></td>
<td><strong>668</strong></td>
<td>--</td>
<td>--</td>
<td><strong>3.137</strong></td>
</tr>
</tbody>
</table>
other factors, as in Dyke and Patterson (1952). The estimates obtained under the two models using the same \( \lambda = .425 \) are presented in Table 2.

Table 2

<table>
<thead>
<tr>
<th>Model 1 (Main Effects Only)</th>
<th>Model 2 (Main Effects and Interactions with Lectures)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>(-.1577 \pm .0928)</td>
</tr>
<tr>
<td>Newspapers (a)</td>
<td>(.2492 \pm .0395)</td>
</tr>
<tr>
<td>Radio (b)</td>
<td>(.1207 \pm .0565)</td>
</tr>
<tr>
<td>Solid Reading (c)</td>
<td>(.4106 \pm .0404)</td>
</tr>
<tr>
<td>Lectures (d)</td>
<td>(.2009 \pm .0966)</td>
</tr>
<tr>
<td>(ab)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(bd)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(cd)</td>
<td>(\ldots)</td>
</tr>
<tr>
<td>(\chi^2)</td>
<td>3.137, 10 d.f.</td>
</tr>
</tbody>
</table>

Thus, we observe that none of the interactions is significant at the 5\% level. However, in a different scale, namely \( \lambda = 0 \), Dyke and Patterson found the interaction (cd) to be significant.

4. Transformation of the Design Variable and the Dose-Response Problem

The general dose-response problem that will be studied occurs when \( k \) groups of subjects are put under experiment. Corresponding to each of the \( k \) groups, there is a particular dosage level to be tested. Let us suppose that \( r_i \) responses are obtained when \( n_i \) subjects are studied at dosage \( x_i \), for \( i = 1, \ldots, k \). The problem then is to fit a cumulative probability distribution to the observed sigmoid response curve. The probability of observing \( r_i \) responses at dosage level \( x_i \) is

\[
P(r_i | x_i) = \left( \frac{n_i}{r_i} \right) \left( \frac{1 - p_i}{1 - p_i} \right) \left( 1 - p_i \right)^{r_i}
\]

where \( p_i = P[Y = 1 | x_i] = P[T < x_i] = G(x_i) \) is the probability of an individual response \((Y = 1)\). Further, it is assumed that \( p_i \) can be expressed in terms of a tolerance distribution \( G \) associated with \( T \).

It is generally understood that a symmetric tolerance distribution will adequately describe the data if a logarithmic transformation of the dosage is used. In fact, most of the published work on this subject has assumed that \( p_i = F(x_i + \beta \log(x_i)) \), where \( F \) is usually either the normal or the logistic distribution. Sometimes, though, experience has suggested a transformation other than log.
In our approach, we follow the suggestion in Cox (1970, p. 110) of applying the Box-Cox transformation to the dosage level. That is to the independent variable. See Box and Tidwell (1962) for a thorough discussion and applications of transformations to independent variables. The aim in the present situation is to nearly symmetrize the original tolerance distribution, even when the assumed model is incorrect.

Thus, we tentatively assume

\[ p_i(\theta_0) = F(\lambda'0)^{\beta_0} - F(\alpha_0 + \beta_0 \lambda_0) \]  

for some parameter values \( \alpha_0, \beta_0 \) and \( \lambda_0 \), where \( F \) is a known symmetric distribution. In fact we consider a location scale family created from a pdf \( f(\cdot) \), which is itself differentiable. The vector parameter \( \theta = (\alpha, \beta, \lambda)^T \) will then be estimated by maximum likelihood. The likelihood function for \( n = \sum n_i \) observations is given by

\[ L(\theta) = \prod_{i=1}^{k} \left( p_i(\theta)^{n_i} (1 - p_i(\theta))^{n_i - r_i} \right) \]

so, the log-likelihood for \( \theta \) is simply

\[ l(\theta) = \sum_{i=1}^{k} \left[ \log(n_i) - \log(n_i - r_i) - \log(r_i) \right] \]

\[ + r_i \log(p_i(\theta)) + (n_i - r_i) \log(1 - p_i(\theta)) \]  

Strong consistency of the MLE is established even when the model (5) is not correct.

**Theorem 1:** For each \( k = 1, \ldots, k \), let \( p_i \) be the probability of a response under the true tolerance distribution \( G \). Let \( p_i(\theta) \) be the probability under the transformed distribution \( F \) and let \( \hat{p}_i \) be the observed frequency of response. If

1. the parameter space \( \theta \subset \mathbb{R}^3 \) is compact,

2. \( \lim_{n \to \infty} u_n \to u \) with \( 0 < u < 1 \) \( \forall \theta \) and \( \sum_{i=1}^{k} u_i = 1 

3. the function \( H(\theta) = \sum_{i=1}^{k} u_i \left( p_i(\theta)^{p_i(\theta)} \right) + \left( 1 - p_i(\theta) \right)^{1 - p_i(\theta)} \) has a unique maximum at

\[ \hat{\theta} = \hat{\theta}_0 = (\hat{\alpha}_0, \hat{\beta}_0, \hat{\lambda}_0)^T \];

then, the MLE \( \hat{\theta}_n \) is a strongly consistent estimator of \( \theta_0 \).

The asymptotic distribution of \( \hat{\theta}_n \) can be obtained based on the asymptotic behavior of the gradient \( V_{\theta} L(\theta) \) and Hessian \( V_{\theta \theta} L(\theta) \) of the log-likelihood. Namely, \( V_{\theta} L(\theta) \) has components

\[ \frac{\partial L(\theta)}{\partial \theta} = \sum_{i=1}^{k} \left( p_i(\theta) \left( 1 - p_i(\theta) \right) \left[ \frac{3p_i(\theta)}{2\theta} \right] \right) u = 1, 2, 3 \]  

and \( V_{\theta \theta} L(\theta) \) has components
\[
\frac{\partial^2 \theta}{\partial \alpha \partial \beta} = \sum_{i=1}^{k} \left\{ \frac{\hat{p}_i(2\hat{p}_i(0) - 1) - p_i(0)^2}{\hat{p}_i(0)(1 - \hat{p}_i(0))^2} \right\} \left( \frac{\partial^2 \hat{p}_i(0)}{\partial \alpha \partial \beta} \right) + \frac{p_i - \hat{p}_i(0)}{\hat{p}_i(0)(1 - \hat{p}_i(0))} \left( \frac{\partial^2 \hat{p}_i(0)}{\partial \alpha^2} \right) + \frac{\partial^2 \hat{p}_i(0)}{\partial \beta^2}
\]
\[
\text{where } \frac{\partial \hat{p}_i(0)}{\partial \alpha} = f(X_i) \frac{\partial}{\partial \alpha} \text{ with } X_i = \alpha + \beta \cdot s_i(k) \forall i.
\]

**Theorem 4:** Let the assumptions (i)-(iv) of Theorem 3 be true and suppose further that

(i) the true parameter value \( \theta_0 \) is an interior point of \( \Omega \)

(ii) \( \sum_{i=1}^{k} \sqrt{\hat{s}_i} \left[ \hat{p}_i(0) \hat{P}_i(0)(1 - \hat{p}_i(0))^2 \right] \hat{P}_i(0) = O \)

(iii) the Hessian of \( \hat{H}(\theta) \), \( \hat{V}^2 \hat{H}(\theta) = \left( \frac{\partial^2 \hat{H}(\theta)}{\partial \theta^2} \right)_{\theta_0} \) is nonsingular at \( \theta_0 \).

Then \( \sqrt{\hat{u}(\theta - \theta_0)} \Rightarrow N_j(0, VV') \) as \( n \rightarrow \infty \), where \( V = \left[ \hat{V}^2 \hat{H}(\theta_0) \right]^{-1} \)

and

\[
\hat{u} = \sum_{i=1}^{k} s_i \left[ \hat{p}_i(0) \hat{P}_i(0)(1 - \hat{p}_i(0))^2 \right] \left[ \hat{P}_i(0) \right]^{1/2} \left[ \hat{P}_i(0) \right]^{1/2}
\]

with \( \hat{V}_i(\theta_0) = \left( \frac{\partial \hat{p}_i(0)}{\partial \theta} \right) \hat{P}_i(0) \).

5. Examples of Transformation with Probit and Logit Models

We first remark that it is perhaps computationally simplest to maximize the log-likelihood function following a two-stage procedure. That is, first fix a value of \( \lambda \) and maximize \( L_\alpha(\alpha, \beta, \lambda) \) over \( \alpha \) and \( \beta \). Then search over values of \( \lambda \). A consistent estimate of the variance-covariance matrix, \( \hat{V} \), is obtained by replacing the true probabilities \( \hat{p}_i(\alpha) \) by the observed frequencies \( n_i \) and the true parameter value \( \hat{\theta}_0 \) by its MLE \( \hat{\theta}_0 \). For illustration, we consider the data shown in Table 3. Finney (1971) analyzed these data to compare the performance of logits vs. logits.

The parameter estimates for the integrated normal model are \( \hat{\alpha} = -71.9201, \hat{\beta} = 51.6027, \hat{\lambda} = -0.587 \), and for the logit model \( \hat{\alpha} = -71.9335, \hat{\beta} = 36.0560, \hat{\lambda} = -0.204 \). The corresponding estimated variance-covariance matrices are:

\[
\begin{pmatrix}
59.5757 \\
-73.1211 & 89.7779 \\
0.5393 & -0.6625 & 0.0493
\end{pmatrix}
\]

and

\[
\begin{pmatrix}
86.3624 \\
-78.5013 & 71.3697 \\
0.8346 & -0.7359 & 0.0811
\end{pmatrix}
\]
Therefore, the age-s/age regression analysis and the chi-square analysis used to evaluate the goodness-of-fit of the model indicate that the model fits the data reasonably well. The chi-square statistic for the model is 12.33, which is less than the critical value of 21.01 at the 0.05 level of significance. This suggests that there is no significant difference between the observed and expected values, indicating that the model is a good fit for the data.
Appendix: Proof of Theorem 4

Let us expand $\mathbb{V}_n \left( \hat{\theta}_n \right)$ in Taylor's series about $\theta_0$, so that

$$\frac{1}{n} \mathbb{V}_n \left( \theta_0 \right) = \frac{1}{n} \mathbb{V}_n \left( \hat{\theta}_n \right) + \frac{1}{n} \mathbb{V}^2_n \left( \theta_0 \right) \left( \mathbb{V}_n \left( \hat{\theta}_n \right) - \theta_0 \right) \text{ for some}$$

$$\theta_n = \frac{\hat{\theta}_n}{n} \left( 1 - \gamma \right) \theta_0, \quad 0 < \gamma < 1.$$  Since $\theta_n \rightarrow \theta_0$ and $\theta_0$ is an interior point of $\hat{\theta}_n$, $\mathbb{V}_n \left( \hat{\theta}_n \right) = 0$ for $n$ sufficiently large, with probability one. Thus, we know that $\frac{1}{n} \mathbb{V}_n \left( \hat{\theta}_n \right)$ and

$$- \frac{1}{n} \mathbb{V}^2_n \left( \hat{\theta}_n \right) \left( \mathbb{V}_n \left( \hat{\theta}_n \right) - \theta_0 \right)$$

have the same limiting distribution.

Next, let us recall that $\gamma_j = \sum_{i=1}^{k} \gamma_{ij}$ where $\gamma_{ij} = 1$ if and only if $Y_j < Y_i$. Therefore, by (6) we get

$$\mathbb{V}_n \left( \theta_0 \right) = \sum_{i=1}^{k} \mathbb{V}_n \left( \theta_0 \right)$$

where

$$\mathbb{V}_n \left( \theta_0 \right) = \sum_{i=1}^{k} \left( Y_i - p_i \left( \theta_0 \right) \right) \left( p_i \left( \theta_0 \right) - 1 - p_i \left( \theta_0 \right) \right)^{-1} \quad \mathbb{V} \left( \theta_0 \right)$$

For each $j$, the random vectors $(Z_{ij} \left( \theta_0 \right), \gamma_{i=1, \ldots, n})$ are iid with

$$\mathbb{E} \left[ Z_{ij} \left( \theta_0 \right) \right] = U_i \quad \text{and} \quad \mathbb{E} \left[ Z_{ij} \left( \theta_0 \right) \right] = \frac{1}{U_i} \quad \text{where}$$

$$U_i = \left( p_i \left( \theta_0 \right) - 1 - p_i \left( \theta_0 \right) \right) \left( p_i \left( \theta_0 \right) - 1 - p_i \left( \theta_0 \right) \right)^{-1} \quad \mathbb{V} \left( \theta_0 \right) \quad \text{and}$$

$$Y_i = \frac{1}{U_i} \left( p_i \left( \theta_0 \right) - 1 - p_i \left( \theta_0 \right) \right) \left( p_i \left( \theta_0 \right) - 1 - p_i \left( \theta_0 \right) \right)^{-1} \quad \mathbb{V} \left( \theta_0 \right) \quad \text{and}$$

$$Y_i = \frac{1}{U_i} \left( p_i \left( \theta_0 \right) - 1 - p_i \left( \theta_0 \right) \right) \left( p_i \left( \theta_0 \right) - 1 - p_i \left( \theta_0 \right) \right)^{-1} \quad \mathbb{V} \left( \theta_0 \right) \quad \text{and}$$

Thus, applying the multivariate CLT $k$ times, we get

$$\frac{1}{n} \mathbb{V}_n \left( \theta_0 \right) \xrightarrow{d} N_{k} \left( 0, I_k \right) \quad \text{as} \quad n \rightarrow \infty, \quad k = 1, \ldots, k.$$

Further,

$$\frac{1}{n} \mathbb{V}_n \left( \theta_0 \right) \xrightarrow{d} N_{k} \left( 0, I_k \right) \quad \text{as} \quad n \rightarrow \infty, \quad k = 1, \ldots, k$$

in such a way that

$$\frac{1}{n} \mathbb{V}^2_n \left( \hat{\theta}_n \right) \left( \mathbb{V}_n \left( \hat{\theta}_n \right) - \theta_0 \right) \xrightarrow{d} N_{k} \left( 0, I_k \right) \quad \text{as} \quad n \rightarrow \infty.$$

with $W$ as given in (8). Now, it can be observed that the second order partial derivatives defined by (7) are uniformly continuous for $0 \leq \theta_0 \leq 1$. Thus, since $p_j$ is a consistent estimator of $p_j$, it follows that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{V}^2_n \left( \hat{\theta}_n \right) \xrightarrow{d} W \left( \theta_0 \right) \quad \text{with probability one and uniformly on} \quad \Omega.$$

Next, as $\frac{1}{n} \mathbb{V}^2_n \left( \hat{\theta}_n \right) \leq \mathbb{V}^2 \left( \theta_0 \right)$, for $0 < \gamma_n < 1$, (10) implies that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{V}^2_n \left( \hat{\theta}_n \right) \xrightarrow{d} W \left( \theta_0 \right) \quad \text{in probability.}$$

Premultiplying (9) by $V = \mathbb{V} \left( \theta_0 \right)^{-1}$ and using Slutsky's Theorem we obtain the desired conclusion.
Bibliography


Dyke, C. V. and Patterson, N. D. (1952). "Analysis of factorial arrangements when the data are proportions." Biometrics 8, 1-12.


