REGRESSION AND ANOVA WITH ZERO-ONE
DATA: MEASURES OF RESIDUAL VARIATION

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 85
APRIL 27, 1976

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MPS74-21416

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
REGRESSION AND ANOVA WITH ZERO-ONE DATA: MEASURES OF RESIDUAL VARIATION

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 85
APRIL 27, 1976

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MPS74-21416

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Regression and ANOVA with zero-one data: measures of residual variation

Bradley Efron

Abstract.

We consider regression situations for which the response variable is dichotomous. The most common analysis fits successively richer linear logistic models, and measures the residual variation from the model by minus twice the maximized log likelihood. General measures of residual variation are considered here, including ordinary squared error and prediction error as well as the log likelihood. All of these are shown to be satisfactory in a certain primitive sense, unlike quantitative regression theory where only squared error is logically satisfactory. The relation of Goodman and Kruskal's measures of categorical association to the Savage-de Finetti theory of probability elicitation is demonstrated.
1. **Introduction.**

Table 1 displays a regression situation in which the response variable takes values either zero or one. "One" represents "tested positive for Toxoplasmosis", "zero" represents "tested negative". There are 697 responses, 356 1's and 341 0's. It happens that the 697 cases are grouped in 34 cities, which conveniently condenses the display of the data, but the methods considered in this paper do not depend on such groupings. A regression analysis of the 0-1 data on the covariate "annual rainfall in the home city of subject" is desired.

**Toxoplasmosis, 34 Cities in El Salvador (Remington et al., 1970)**

<table>
<thead>
<tr>
<th>City #</th>
<th># Pos.</th>
<th># Samp.</th>
<th>Annual Rainfall in mm.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1/4</td>
<td>3/10</td>
<td>1/5</td>
</tr>
<tr>
<td></td>
<td>3/10</td>
<td>2/2</td>
<td>3/5</td>
</tr>
<tr>
<td></td>
<td>2/8</td>
<td>7/19</td>
<td>3/6</td>
</tr>
<tr>
<td></td>
<td>1735</td>
<td>1936</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>1973</td>
<td>1750</td>
<td>1800</td>
</tr>
<tr>
<td></td>
<td>1750</td>
<td>2077</td>
<td>1920</td>
</tr>
<tr>
<td>10</td>
<td>8/10</td>
<td>7/24</td>
<td>0/1</td>
</tr>
<tr>
<td></td>
<td>15/30</td>
<td>4/22</td>
<td>0/1</td>
</tr>
<tr>
<td></td>
<td>6/11</td>
<td>0/1</td>
<td>33/54</td>
</tr>
<tr>
<td></td>
<td>4/9</td>
<td>5/18</td>
<td>2/12</td>
</tr>
<tr>
<td></td>
<td>1800</td>
<td>2050</td>
<td>1630</td>
</tr>
<tr>
<td></td>
<td>1650</td>
<td>2200</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>1770</td>
<td>1920</td>
<td>1770</td>
</tr>
<tr>
<td></td>
<td>2240</td>
<td>1620</td>
<td>1756</td>
</tr>
<tr>
<td>11</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>32</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>TOTAL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0/1</td>
<td>8/11</td>
<td>41/77</td>
</tr>
<tr>
<td></td>
<td>24/51</td>
<td>7/16</td>
<td>46/82</td>
</tr>
<tr>
<td></td>
<td>9/13</td>
<td>23/43</td>
<td>52/75</td>
</tr>
<tr>
<td></td>
<td>8/13</td>
<td>3/10</td>
<td>1/6</td>
</tr>
<tr>
<td></td>
<td>5/11</td>
<td>356/697</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1650</td>
<td>2250</td>
<td>1796</td>
</tr>
<tr>
<td></td>
<td>1890</td>
<td>1871</td>
<td>2063</td>
</tr>
<tr>
<td></td>
<td>2100</td>
<td>1918</td>
<td>1834</td>
</tr>
<tr>
<td></td>
<td>1780</td>
<td>1900</td>
<td>1976</td>
</tr>
<tr>
<td></td>
<td>2292</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Number of subjects testing positive for Toxoplasmosis in 34 cities of El Salvador, versus number sampled. All subjects are between 11 and 15 years old. Abstracted from a larger data set in Remington et al., [20].
Table 2 presents the most common such analysis. The logit for each of the 697 cases, \( \log P(1)/P(0) \), is fitted to the data by an increasing sequence of linear models. The fit at each stage is by maximum likelihood. The column "-2 log \( f^* \)" is minus twice the log of the maximized density. This quantity is treated as a measure of residual variation, analogous to residual squared error in a standard regression analysis. Differences of successive values of \(-2 \log f^*\) give a standard-looking ANOVA table, the entries of the "SS" column being Wilks' maximum likelihood ratio statistic for testing an hypothesis versus a more general alternative. See Cox [4] and Duncan and Walker [6].

Table 2 raises several interesting questions. The cubic term is quite significant by Wilks' criterion, but not by the approximate \( F \) test which compares this quantity to the corresponding term for Groups. Which is correct? Is the last "Mean Square", which in the ordinary situation serves as an estimate of essential error, of any use here? Most importantly, is \(-2 \log f^*\) really the correct analogue of the residual sum of squares for quantitative regression theory?
Table 2. A logistic regression of the Toxoplasmosis data. Increasing models for log $P(1)/P(0)$ are fitted as functions of rainfall, $z$. Minus twice the maximized log likelihood is used as a measure of residual variation to construct the ANOVA-like table. The cubic term is significant by Wilks' criterion, but not by the analogue of the $F$ test.
The last question is the main topic considered here. We introduce a general class of residual variation measures for 0-1 data. Besides $-2 \log f^*$ this class includes ordinary squared error and prediction error, the number of predicted values of $P(1)$ on the opposite side of 1/2 from the corresponding data point. All of these measures are shown to satisfy the one-way ANOVA table and therefore in a certain minimal sense to be satisfactory as variation measures.

Goodman and Kruskal [11,12,13,14], Light and Margolin [17,18], and many other writers have considered residual variation measures for categorical data in the special case of the one-way layout. One objective of our paper is to show the close connection between this work and the Savage-deFinetti[21,5] theory of subjective probability elicitation.

Efficiency considerations for estimation and hypothesis testing do lead directly to the use of $-2 \log f^*$. This theory, which is mostly familiar, is sketched briefly in the last two sections. Without attempting a deep or detailed analysis of the asymptotics involved we present some remarks on the interpretation of ANOVA-like tables such as that for the Toxoplasmosis data.

From the point of view of logical neatness it is too bad that no one measure of residual variation for binary data is uniquely best. On the other hand Goodman and Kruskal emphasize the opportunity this gives the statistician to fashion measures particularly apt for the problem at hand. My own experience has been that it is often useful to print out all three major measures, $-2 \log f^*$, squared error, and prediction error, and adds virtually nothing to the computational burden.
2. **Measures of Binary Variation.**

Standard ANOVA theory is based on a measure of variation between an observed data point \( x \) and an explanatory point \( \mu \), namely

\[
S(x, \mu) = (x - \mu)^2.
\]  

(2.1)

If \( \mathbf{x} = (x_1, x_2, \ldots, x_n) \) represents all the data, and \( \mathbf{\mu} = (\mu_1, \mu_2, \ldots, \mu_n) \) is a vector of explanations, then the total residual variation of the explanation is

\[
S(\mathbf{x}, \mathbf{\mu}) = \frac{1}{n} \sum_{i=1}^{n} S(x_i, \mu_i) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_i)^2.
\]

(2.2)

The possible explanatory vectors \( \mathbf{\mu} \) are indexed by a lower dimensional vector \( \mathbf{\hat{\theta}} \), usually in a linear manner, and the theory proceeds by minimizing \( S(\mathbf{x}, \mathbf{\mu}(\mathbf{\hat{\theta}})) \) over successively richer models.

In the binary case the observed random variables only take on values 0 or 1. Let \( \pi \) indicate a possible explanatory value for such a variate, \( 0 \leq \pi \leq 1 \), and for convenience denote

\[
x = 1 - \pi, \quad s(\pi) = S(1, \pi).
\]

(2.3)

We consider measures of variation \( S(x, \pi) \) satisfying two elementary properties,

\[
S(0, \pi) = S(1, X) = s(X),
\]

(2.4)

and

\[
s(\pi) \text{ nonincreasing in } \pi, \quad s(1) = 0.
\]

(2.5)
The first of these is a symmetry property while the second says that
$S(x, \pi)$ gets smaller as $\pi$ gets closer to $x$. A third less obvious
restriction on the form of $S(x, \pi)$ is added at (2.12).

Now let $\xi = (x_1, x_2, \ldots, x_n)$ be a vector of 0's and 1's, and
$\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ be a vector of explanation, $0 \leq \pi_i \leq 1$ for
$i = 1, 2, \ldots, n$. The total variation of $\xi$ from $\pi$ is defined to be

$$S(\xi, \pi) = \sum_{i=1}^{n} S(x_i, \pi_i),$$

(2.6)

In particular let $\xi = n\pi = (\pi, \pi, \ldots, \pi)$. The total variation is

$$S(\xi, \pi_l) = n[ps(\pi) + qs(x)]$$

(2.7)

where

$$p = \frac{\#(x_i = 1)}{n}, \quad q = 1 - p = \frac{\#(x_i = 0)}{n}.$$  

(2.8)

We define the functions $\sigma(p, \pi)$ and $\sigma(p)$, $0 \leq p, \pi \leq 1$, to be

$$\sigma(p, \pi) = ps(\pi) + qs(x)$$

(2.9)

and

$$\sigma(p) = \sigma(p, p) = ps(p) + qs(q).$$

(2.10)

In other words $\sigma(p, \pi)$ is the average variation of the $n$ values
$x_1, x_2, \ldots, x_n$ about the single explanatory point $\pi$. If $X$ is binomially
distributed with one repetition and probability of success $p$, $X \sim Bi(1, p)$, then
\[
\sigma(p, \pi) = ES(X, \pi),
\]  
(2.11)

which can be thought of as definition (2.9) with \( n \to \infty \).

The additional requirement we make on \( S(x, \pi) \) is that

\[
\sigma(p) = \sigma(p, p) = \min_{0 \leq \pi \leq 1} \sigma(p, \pi). 
\]  
(2.12)

This says that the best single number \( \pi \) for describing a vector \( X \)
of \( np \) 1's and \( nq \) 0's is \( \pi = p \), in the sense that this choice
minimizes the total residual variation \( S(x, \pi_1) \). Equation (2.12) shows
that \( \sigma(\cdot) \) is concave since it is the minimum of linear functions.

Figure 1 illustrates the situation. For each value of \( \pi \) between
0 and 1, \( s(\pi) \) is plotted along the right vertical axis and \( s(x) = s(1-\pi) \)
along the left vertical axis. Then \( \sigma(p, \pi) = ps(\pi) + qs(x) \) is the straight
line connecting these two points. Condition (2.12) says that the line
corresponding to \( \pi = p \) is tangent to the concave function \( \sigma(\cdot) \) at \( p \).

If \( s(\cdot) \) is differentiable then (2.9) gives

\[
\frac{\partial \sigma(p, \pi)}{\partial \pi} = ps'(\pi) - qs'(x), 
\]  
(2.13)

so setting the derivative equal to zero at \( \pi = p \) gives a necessary
condition for (2.12),

\[
ps'(p) = qs'(q). 
\]  
(2.14)

It will turn out in Section 4 that this is also sufficient. Subject only
to the monotonicity requirement (2.5), the function \( s(\cdot) \) can be defined
Figure 1. The basic quantities involved in the measurement of binary variation. The function \( S(1, \pi) = s(\pi) \) measures the variation of the observed value \( x=1 \) from the explanatory point \( \pi \); \( n\sigma(p, \pi) \) is the variation of \( np \) 1's and \( nq \) 0's about the constant value \( \pi \). The concave function \( n\sigma(p) \) measures the internal variation of \( np \) 1's and \( nq \) 0's. The difference \( S(p, \pi) = \sigma(p, \pi) - \sigma(p) \) is defined to be the variation of explanation \( p \) from explanation \( \pi \).

arbitrarily on \( [\frac{1}{2}, 1] \), but is then determined on \( [0, \frac{1}{2}] \) by means of (2.14). See example 4, section 3.

The function \( \sigma(p) \) is a measure of internal variation. More exactly, \( n\sigma(p) \) is the minimum possible variation of \( np \) 1's and \( nq \) 0's about any single central point. It is reasonable to require \( \sigma(p) = \sigma(q) \) and \( \sigma(0) = 0 \), though not actually necessary to develop
the theory. Concavity is also a reasonable requirement. Suppose we have two groups of 0-1 data of sizes $n_1$ and $n_2$, say $\xi^{(1)}$ and $\xi^{(2)}$, the proportions of 1's being $p_1$ and $p_2$. Then

$$p = (n_1p_1 + n_2p_2)/(n_1 + n_2)$$

is the proportion of 1's in the combined data set. The concavity relationship

$$n_1\sigma(p_1) + n_2\sigma(p_2) \leq (n_1 + n_2)\sigma(p)$$

is equivalent in our notation to

$$S(\xi^{(1)}, p_1) + S(\xi^{(2)}, p_2) \leq S((\xi^{(1)}, \xi^{(2)}), p), \quad (2.16)$$

(the $\xi$'s being of correct dimension). This just says that we can explain the data set $(\xi^{(1)}, \xi^{(2)})$ better with a separate central value $p_1$ for each group than with a single central value $p$ for all $n_1 + n_2$ numbers.

Figure 1 shows that starting with any concave function $\sigma(\cdot)$, symmetric about $1/2$ with $\sigma(0) = 0$, we can define a function $s(\cdot)$ satisfying (2.5) and (2.12). The definition is unique if $\sigma(\cdot)$ is differentiable. The theory developed in the following sections depends only on the shape of the function $\sigma(\cdot)$. In this sense assumption (2.12) is innocuous once we have agreed that (2.16) is necessary.

Savage [21] and de Finetti[5] have developed a theory for eliciting subjective probabilities which is formally identical to figure 1. Suppose a forecaster has to assign a probability $\pi$ to the event that a binary random variable $X$ takes value 1 rather than 0. Let $S(x, \pi)$ for $x = 0, 1$ be the penalty he pays if it is subsequently
observed that \( X = x \). If actually the forecaster believes \( P(X=1) = p \) then his expected penalty for stating \( P(X=1) = \pi \) is perceived to be 
\[
\sigma(p, \pi) = pS(1, \pi) + qS(0, \pi).
\]
In order to make the forecaster express his true opinion about \( P(X=1) \) it is necessary that \( \sigma(p, p) = \min_{\pi} \sigma(p, \pi) \) as at (2.12).

Figure 1 shows how the definition of \( S(\cdot, \pi) \) naturally extends to values of the first argument anywhere in the interval \([0,1]\),
\[
S(p, \pi) \equiv \sigma(p, \pi) - \sigma(p).
\]  \( (2.17) \)

From (2.17) and (2.10) we see that
\[
nS(p, \pi) = \bar{S}(\bar{\chi}, \bar{\pi}_1) - \bar{S}(\bar{\chi}, p_1),
\]  \( (2.18) \)

the excess variation of a vector of \( np \) 1's and \( nq \) 0's about \( \pi \) rather than the minimizing value \( p \). By component-wise addition (2.17) extends to a measure of variation between two explanatory vectors \( \bar{\chi} \) and \( \bar{\pi} \) with components in \([0,1]\),
\[
\bar{S}(\bar{\chi}, \bar{\pi}) \equiv \sum_{i=1}^{n} S(p_i, \pi_i).
\]

In the ANOVA theory of sections 5 and 7, \( \bar{\pi} \) is chosen from a wider class of possible explanations than \( \pi \). The extreme case is when \( \bar{\pi} \) can be anything, in which case it can be chosen equal to the observed data \( \bar{\chi} \), getting us back to \( \bar{S}(\bar{\chi}, \bar{\pi}) \).

Definition (2.6) deserves some discussion. It is simple and pleasing to make the total variation \( \bar{S}(\bar{\chi}, \bar{\pi}) \) the sum of the component variations \( S(x_i, \pi_i) \). However, this excludes certain obvious measures.
of binomial variation such as

$$\sum_{i=1}^{I} \frac{n_i (p_i - \pi_i)^2}{\pi_i X_i}$$  \hspace{1cm} (2.19)

(where the binary variables are observed in I groups, \(n_i\) in the \(i^{th}\) group of which \(n_i p_i\) are 1's). See the discussion in sections 3 and 4. Grouped measures such as (2.19) suffer from the practical difficulties of grouping when there is enough covariate information to make each case unique. Theoretically they are unpleasant in not being additive when separate groups of data are combined.

3. Examples.

Example 1. Entropy. Let \(S(1, \pi) = S(0, X) = s(\pi)\) be given by

$$s_1(\pi) \equiv -2 \log \pi$$  \hspace{1cm} (3.1)

so that

$$\sigma_1(p, \pi) = -2[p \log p + q \log q],$$  \hspace{1cm} (3.2)

$$\sigma_1(p) = -2[p \log p + q \log q],$$

and

$$S_1(p, \pi) = 2[p \log p/\pi + q \log q/\pi].$$  \hspace{1cm} (3.3)

The function \(\sigma_1(p)\), or more exactly \(c\sigma_1(p)\) where \(c = (2 \log 2)^{-1} = .731\), is called the "entropy"; \(S_1(p, \pi)/2\) is the Kullback-Leibler distance from \(p\) to \(\pi\).
We can write \( S_1(x, \pi) = -2[x \log \pi + (1-x) \log \lambda] \), using the convention \( 0 \log 0 = 0 \), so for \( \tilde{x} = (x_1, x_2, \ldots, x_n) \) and \( \tilde{\pi} = (\pi_1, \pi_2, \ldots, \pi_n) \),
\[
S_1(\tilde{x}, \tilde{\pi}) = -2 \sum_{i=1}^{n} \log(\pi_i x_i^{1-x_i}) \tag{3.4}
\]
\[
= -2 \log f_{\tilde{x}}(\tilde{x}),
\]
where \( f_{\tilde{x}}(\tilde{x}) \) is the likelihood of \( \tilde{x} \) when \( X_1 \sim \text{Bi}(1, \pi_i) \) independently for \( i = 1, 2, \ldots, n \). Definition (3.1) leads to the analysis in table 2. The factor "2" is included in deference to the usual statement of Wilks' theorem, see section 7.

The quantity \( S_1(\tilde{x}, \tilde{\pi}) \) has an information-theoretic interpretation. If \( \tilde{x} \) has density \( f_{\tilde{x}}(\tilde{x}) \) then an optimal code will require very nearly \( c \sum_{i=1}^{n} \sigma_1(\pi_i) \) bits on the average to transmit \( \tilde{x} \), see [22]. For large \( n \) we expect \( S_1(\tilde{x}, \tilde{\pi}) \) to be near its expectation \( c \sum_{i=1}^{n} \sigma_1(\pi_i) \), so that the residual variation \( cS_1(\tilde{x}, \tilde{\pi}) \) approximately measures how many bits are necessary to transmit \( \tilde{x} \) given knowledge of \( \tilde{\pi} \).

Example 2. Squared Error. Define
\[
s_2(\pi) = 6\pi^2, \tag{3.5}
\]
so
\[
\sigma_2(p, \pi) = 6[p \pi^2 + q \pi^2] \tag{3.6}
\]
\[
\sigma_2(p) = 6pq,
\]
and
\[
S_2(p, \pi) = 6(p-\pi)^2, \quad S_2(\theta, \pi) = 6||\theta - \pi||^2. \tag{3.7}
\]
The factor "6" is explained in section 4. Analysis based on $s_2(\cdot)$ is equivalent to ordinary unweighted regression or ANOVA. The measure of internal variation $\sigma_2(p)$ is proportional to the binomial variance npq. Among all choices of $s(\cdot)$ satisfying (2.5) and (2.12), only $s(\pi) = c\chi^2$ makes $S(p,\pi)$ a function of $p-\pi$, and likewise only these make $S(p,\pi) = S(\pi,p)$. See Savage [21]. Many authors have advocated the use of $S_2(p,\pi)$ in constructing ANOVA tables for 0-1 data, see in particular Gini [10] and Light and Margolin [17].

Example 3. Prediction Error. Define

$$s_3(\pi) = \begin{cases} 0 & \text{if } \pi \geq \frac{1}{2} \\ 1 & \text{if } \pi < \frac{1}{2}. \end{cases} \quad (3.8)$$

Then

$$\sigma_3(p,\pi) = \begin{cases} q & \text{if } \pi \geq \frac{1}{2} \\ \sigma_3(p) = \min\{p,q\}, & \sigma_3(p) = \min\{p,q\}, \\ p & \text{if } \pi < \frac{1}{2}. \end{cases} \quad (3.9)$$

and

$$S_3(p,\pi) = \begin{cases} |p-q| & \text{if } \pi < \frac{1}{2}, \ p > \frac{1}{2} \text{ or } \pi > \frac{1}{2}, \ p < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \quad (3.10)$$

Notice that
\[ S_2(x, \pi) = \#\{(x_1, \pi_1): \pi_1 \leq \frac{1}{2}, x_1 = 1 \text{ or } \pi_1 \geq \frac{1}{2}, x = 0\} . \] (3.11)

The total variation \( S_2(x, \pi) \) is the number of predicted values \( \pi_1 \) on the wrong side of \( 1/2 \), which is the usual definition of prediction error.

![Graph showing \( \sigma(p) \) for different functions]  

Figure 2. The functions \( \sigma(\cdot) \) for 1) entropy, 2) squared error, and 3) prediction error.

Example 4. Linear Error. Suppose we wish the measure of variation to increase linearly with the error of prediction, \( S_4(1, \pi) = s_4(\pi) = x \). By (2.14), \( s_4(\pi) = x \) for \( \pi \geq \frac{1}{2} \) implies that \( s_4(\pi) = \pi - \log 2\pi \) for \( \pi < \frac{1}{2} \) (assuming \( s_4(\pi) \) is continuous at \( \pi = \frac{1}{2} \)). We can't use \( s(\pi) = x \) for all values of \( \pi \) and still satisfy condition (2.12).
The measure of internal variation corresponding to \( s'_4(\cdot) \) is
\[
\sigma_4(p) = q \log(e/2q) \text{ for } p \geq \frac{1}{2}.
\]

Every symmetric concave function \( \sigma(\cdot) \) generates a variation function \( S(p, \pi) \) by means of the tangency construction of Figure 1. However, some commonly used variation functions cannot be so constructed. Three such non-examples, as proved in section 4, are \( S(p, \pi) = (p-\pi)^2 / \pi X \), \( S(p, \pi) = (p-\pi)^2 / pq \), and \( S(p, \pi) = \pi \log \pi / p + X \log X / q \). The first two of these relate to chi-squared methods, while the third is Kullback's [16] "minimum discrimination information".

Given any two \( \sigma(\cdot) \) functions, say \( \sigma_A(p) \) and \( \sigma_B(p) \), we can construct the linear combination \( \sigma_{a,b}(p) = a\sigma_A(p) + b\sigma_B(p) \) for \( a \) and \( b \) arbitrary position constants, and use this as a measure of internal variation. It is easy to verify that
\[
s_{a,b}(\pi) = as_A(\pi) + bs_B(\pi), \quad S_{a,b}(p, \pi) = aS_A(p, \pi) + bS_B(p, \pi), \quad \sigma_{a,b}(p, \pi) = a\sigma_A(p, \pi) + b\sigma_B(p, \pi).
\]
This gives a simple way of compromising between different measures of residual variation. For example \( \sigma_1(p) + 4\sigma_3(p) \) represents an approximately equal compromise (see section 4) between entropy and prediction error.

4. **Elementary Relationships.**

Some simple but useful relationships exist between the functions \( s(\pi), \sigma(p, \pi), \sigma(p) \) and \( S(p, \pi) \). For convenient reference we recall from section 2 that
\[
\sigma(p, \pi) = ps(\pi) + qs(X), \quad (4.1)
\]

15
\[ \sigma(p) = ps(p) + qs(q), \quad (4.2) \]

\[ S(p, \pi) = \sigma(p, \pi) - \sigma(p) = p(s(\pi) - s(p)) + q(s(\pi) - s(q)), \quad (4.3) \]

\[ \frac{\partial \sigma(p, \pi)}{\partial \pi} = ps'(\pi) - qs'(\chi), \quad (4.4) \]

\[ ps'(p) = qs'(q). \quad (4.5) \]

The last two of these assume \( s(\cdot) \) differentiable, an assumption continued in this section except in Lemma 2. Differentiating \((4.2)\), and using \((4.5)\), gives

\[ \sigma'(p) = s(p) - s(q) \quad (4.6) \]

and

\[ \sigma''(p) = s'(p) + s'(q) = \frac{s'(p)}{q}. \quad (4.7) \]

Assumption \((2.5)\), that \( s(p) \) is nonincreasing in \( p \), then shows that \( \sigma''(p) \leq 0 \), so \( \sigma(\cdot) \) must be a concave function. We now have verified \((2.12)\), \( \sigma(p, \pi) = \min_{\pi} \sigma(p, \pi) \), starting from definition \((4.1)\), \((2.5)\), and the local minimality condition \((4.5)\).

Differentiating \((4.1)\) and \((4.3)\) gives

\[ \frac{\partial \sigma(p, \pi)}{\partial p} = s(\pi) - s(\chi), \quad \frac{\partial S(p, \pi)}{\partial p} = [s(\pi) - s(p)] - [s(\chi) - s(q)] \quad (4.8) \]

and

\[ \frac{\partial \sigma(p, \pi)}{\partial \pi} = \frac{\partial S(p, \pi)}{\partial \pi} = \frac{s'(p)(p-\pi)}{\chi}. \quad (4.9) \]
(4.9) implies that \( \frac{\partial S(p, \pi)}{\partial \pi} / (p-\pi) \) is not a function of \( p \). This provides a convenient necessary condition for testing whether a given \( S(p, \pi) \) can actually be generated by the tangency construction of Figure 1. Taking \( S(p, \pi) = \pi \log \pi/p + X \log X/q \) for example gives 
\[
\frac{\partial S(p, \pi)}{\partial \pi} / (p-\pi) = (\log \pi q / X p) / (p-\pi),
\]
a function of \( p \), so no choice of \( \sigma(p) \) results in this \( S(p, \pi) \) function. Both \( S(p, \pi) = (p-\pi)^2 / \pi X \) and \( S(p, \pi) = (p-\pi)^2 / pq \) are ruled out on the same grounds.

Second derivatives of \( S(p, \pi) \) and \( \sigma(p, \pi) \) can be calculated assuming that \( s''(\cdot) \) exists. Differentiating (4.9) gives
\[
\frac{\partial^2 \sigma(p, \pi)}{\partial \pi^2} = \frac{\partial^2 S(p, \pi)}{\partial \pi^2} \tag{4.10}
\]
\[
= s''(\pi) \frac{(p-\pi)}{\chi} - s'(\pi) \frac{\sigma}{\chi^2} .
\]

At \( \pi = p \) we get
\[
\left. \frac{\partial^2 S(p, \pi)}{\partial \pi^2} \right|_{\pi=p} = \left. \frac{\partial^2 \sigma(p, \pi)}{\partial \pi^2} \right|_{\pi=p} \tag{4.11}
\]
\[
= - \frac{s'(p)}{q} = - \sigma''(p) = \frac{\partial^2 S(p, \pi)}{\partial p^2} .
\]

Figure 1 shows that
\[
S(p) = \sigma(p) + q \sigma'(p) , \tag{4.12}
\]
\[
S(q) = \sigma(p) - p \sigma'(p) ,
\]
resulting in the following lemma.
Lemma 1.

\[ \int_0^1 \int_0^1 S(p, \pi) d\pi dp = \int_0^1 \sigma(p) dp = \frac{1}{2} \int_0^1 s(\pi) d\pi. \quad (4.13) \]

Proof. Let \( c = \int_0^1 \sigma(p) dp = -\int_0^1 p \sigma'(p) dp \) (by parts), so

\[ 2c = \int_0^1 [\sigma(p) - p \sigma'(p)] dp = \int_0^1 s(q) dp = \int_0^1 s(p) dp, \quad (4.14) \]

which is the last equality in (4.13). Also, since \( S(p, \pi) = ps(\pi) + qs(\pi) \sigma(p) \),

\[ \int_0^1 \int_0^1 S(p, \pi) d\pi dp = \int_0^1 \int_0^1 [ps(\pi) + qs(\pi) \sigma(p)] d\pi dp \]

\[ = 2c - c = c. \quad (4.15) \]

Lemma 1 shows that if we scale \( \sigma(p) \) so it integrates to one, or equivalently scale \( s(\pi) \) so it integrates to 2, then randomly choosing \( p \) and \( \pi \) from \([0,1]\) results in \( S(p,\pi) \) having expected value one.

Example 1 of section 3 satisfies \( \int_0^1 \sigma_1(p) dp = 1 \). The choice of the constant "6" in example 2 makes \( \int_0^1 \sigma_2(p) dp = 1 \). In most data sets investigated by the author, \( S_1(p,\pi) \) and \( S_2(p,\pi) \) give reasonably similar results, as lemma 1 suggests. (See the penicillin example in section 7.) Multiplying \( \sigma_3(\cdot) \) by \( \frac{4}{3} \) gives integral 1, but this destroys the natural interpretation of \( S_3(x,\pi) \) as the number of prediction errors.
Now let "\( \mathbb{E} \)" stand for expectation with respect to some arbitrary probability distribution on a random variable \( P \) with mean value \( p = \mathbb{E}P \).

**Lemma 2.** For any value of \( \pi \) in \([0,1]\),

\[
\mathbb{E}S(P,\pi) = S(p,\pi) + \mathbb{E}S(P,p) .
\]  \hspace{1cm} (4.16)

**Proof.** \( S(P,\pi) - S(P,p) = \sigma(P,\pi) - \sigma(P,p) \) by (4.3), so

\[
\mathbb{E}S(P,\pi) - \mathbb{E}S(P,p) = \mathbb{E}\sigma(P,\pi) - \mathbb{E}\sigma(P,p)
\]  \hspace{1cm} (4.17)

\[
= \sigma(p,\pi) - \sigma(p,p) = S(p,\pi) ,
\]

using (4.1) which says that \( \sigma(p,\pi) \) is linear in \( p \). (Notice that \( s(.) \) is not assumed differentiable here.)

In the case of squared error lemma 2 gives the familiar result

\[
\mathbb{E}(P-\pi)^2 = (p-\pi)^2 + \mathbb{E}(p-p)^2 .
\]  \hspace{1cm} (4.18)

The fact that the same relationship holds for any \( S(p,\pi) \) function obtained from the tangency construction plays an important role in section 5.

The function \( S(p,\pi) \) is the excess of \( \sigma(p,\pi) = \mathbb{E}S(X,\pi) \) over \( \sigma(p,p) = \mathbb{E}S(X,p) \) for \( X \sim \text{Bi}(1,p) \). We can iterate this construction. Using the notation of lemma 2, let
\[ \tilde{\sigma}(p, \pi) = \epsilon S(p, \pi), \]
\[ \tilde{\sigma}(p) = \min_{\pi} \tilde{\sigma}(p, \pi), \quad (4.19) \]
\[ \tilde{S}(p, \pi) = \tilde{\sigma}(p, \pi) - \tilde{\sigma}(p). \]

It turns out that functions \( S(\cdot, \cdot) \) satisfying (2.4), (2.5), (2.12) are "reproducing" in the sense that
\[ \tilde{S}(p, \pi) = S(p, \pi) \quad (4.20) \]
for all values of \( p \) and \( \pi \) in [0,1]. To prove this notice that

lemma 2 is equivalent to
\[ \tilde{\sigma}(p, \pi) = S(p, \pi) + \tilde{\sigma}(p, p). \quad (4.21) \]

Since \( \min_{\pi} S(p, \pi) = S(p, p) = 0 \),
\[ \tilde{\sigma}(p) = \min_{\pi} \tilde{\sigma}(p, \pi) = \tilde{\sigma}(p, p) \quad (4.22) \]
as at (2.12). Then (4.21) and definitions (4.19) give (4.20).

5. **One Way ANOVA.**

Suppose that the experimental units are grouped together in some way, for instance by cities as in the Toxoplasmosis example. Let \( x_{ij} \) be the \( j^{th} \) response in the \( i^{th} \) group, \( i = 1,2,\ldots,I, \ j = 1,2,\ldots,n_i \), and let
\[ p_i = \sum_{j=1}^{n_i} x_{ij}/n_i \quad (5.1) \]
be the proportion of 1's in group \( i \). Also define
\begin{align}
    n &= \sum_{i=1}^{I} n_i, \quad p &= \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} x_{i,j}}{n},
\end{align}

so \( p \) is the overall proportion of 1's in all \( n \) responses. In the Toxoplasmosis example, \( I = 34, n = 697, p_1 = 2/4, p_2 = 3/10, \ldots, p_{34} = 5/11 \) and \( p = 356/697 \).

The vectors
\begin{align}
    \mathbf{x} &= (x_{11}, x_{12}, \ldots, x_{1n_1}, x_{21}, x_{22}, \ldots, x_{2n_2}, \ldots, x_{I1}, x_{I2}, \ldots, x_{In_I}) \\
    \mathbf{p} &= (p_1, p_1, \ldots, p_1, p_2, p_2, \ldots, p_2, \ldots, p_I, p_I, \ldots, p_I) \\
    \mathbf{p}_1 &= (p, p, p, \ldots, p, p, p, \ldots, p, p, p, \ldots, p) \\
    \mathbf{p}_\infty &= (\pi, \pi, \pi, \ldots, \pi, \pi, \pi, \ldots, \pi, \pi, \pi, \ldots, \pi)
\end{align}

respectively represent the data, the best explanation of the data fitting a separate constant to each group (in the sense of minimizing residual variation), the best explanation fitting a single overall constant, and an arbitrary preconceived constant explanation.

The standard one-way ANOVA table is obtained by differencing the residual variation of \( \mathbf{x} \) for \( \mathbf{p}_1, \mathbf{p}_2, \) and \( \mathbf{p}_\infty \). For example
\[ S(\mathbf{x}, \mathbf{p}_1) - S(\mathbf{x}, \mathbf{p}_\infty) \]
is the usual "Between groups sum of squares". Pythagoras' theorem shows that for squared error variation this quantity equals \( S(\mathbf{p}, \mathbf{p}_1) \), the variation of \( \mathbf{p} \) from \( \mathbf{p}_1 \). The gist of theorem 1 is that these Pythagorean relationships hold for any of the measures developed in section 2.
<table>
<thead>
<tr>
<th>Model</th>
<th>Residual Variation</th>
<th>SOURCE</th>
<th>&quot;SS&quot; (difference in residual variation)</th>
<th>degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arbitrary Constant $\pi$</td>
<td>$S(x, \pi)$</td>
<td>MEAN</td>
<td>$S(p_l, \pi)$ = $nS(p, \pi)$</td>
<td>1</td>
</tr>
<tr>
<td>Best Constant $p$</td>
<td>$S(x, p_l)$</td>
<td>BETWEEN GROUPS</td>
<td>$S(p, p_l)$ = $\sum_{i=1}^{I} n_i S(p_i, p)$</td>
<td>$I-1$</td>
</tr>
<tr>
<td>Best Constant $p_i$ each group</td>
<td>$S(x, p)$</td>
<td>WITHIN GROUPS</td>
<td>$S(x, p_l)$ = $\sum_{i=1}^{I} n_i \sigma(p_i)$</td>
<td>$n-1$</td>
</tr>
<tr>
<td>Best Constant each data point</td>
<td>0</td>
<td>TOTAL</td>
<td>$S(x, \pi_l)$ = $n\sigma(p, \pi)$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

Table 3. The standard one-way ANOVA table is obtained by successively differencing the residual variation. Theorem 1 shows that the expressions for these differences in the "SS" column are valid for any of the measures of variation introduced in section 2.

**Theorem 1.** Let $S(p, \pi)$ be any measure of variation (2.17), as illustrated in figure 1. Then

$$S(x, p_l) = S(x, p) + S(p, p_l) ,$$  \hspace{1cm} (5.4)

$$S(p, \pi_l) = S(p, p_l) + S(p_l, \pi_l) ,$$  \hspace{1cm} (5.5)

and

$$S(x, \pi_l) = S(x, p) + S(p, p_l) + S(p_l, \pi_l) .$$  \hspace{1cm} (5.6)
Proof. Relationship (4.3), \( \sigma(x_i, p) = \sigma(p_i) + S(p_i, p) \) implies

\[
\frac{1}{n_1} \sum_{j=1}^{n_1} S(x_{ij}, p) = S(p_1, p) + \frac{1}{n_1} \sum_{j=1}^{n_1} S(x_{ij}, p_i) . \tag{5.7}
\]

Multiplying (5.7) by \( n_1 \) and summing over \( i \) gives (5.4). Changing \( p \) to \( \pi \) in (5.4) gives

\[
S(x, \pi) = S(x, p) + S(p, \pi) . \tag{5.8}
\]

In lemma 2 let \( "E" \) represent a distribution putting mass \( n_1/n \) at \( p_i \), \( i = 1, 2, \ldots, I \), giving

\[
\sum_{i=1}^{I} \frac{n_i}{n} S(p_i, \pi) = S(p, \pi) + \sum_{i=1}^{I} \frac{n_i}{n} S(p_i, p) . \tag{5.9}
\]

Multiplying (5.9) by \( n \) gives (5.5). Substituting (5.5) into (5.8) gives (5.6).

The "coefficient of determination"

\[
R^2 = \frac{S(x, p_1) - S(x, p)}{S(x, p_1)} = \frac{S(p, p_1)}{S(x, p_1)} \tag{5.10}
\]

is the proportional decrease in residual variation obtained in going from the crude explanation \( p_1 \) to the more accurate explanation \( p \). The second equality in (5.10) shows that \( R^2 \) is also the variation of \( \hat{x} \) from \( p_1 \) compared to the variation of \( \hat{x} \) from \( \hat{p} \). \( R^2 = .70 \) for instance means both that the explanation is 70\% as big as the original variation and that the residual variation is 30\% as big as the original
variation. Interpreting $R^2$ would be considerably more difficult if both were not true. This familiar property of squared error holds for all the measures considered here because of (5.4). In the author's opinion (5.4) is the very least we should expect of any reasonable measure of residual variation.

Goodman and Kruskal's [11] measure of association $\tau$, is $R^2$ for $S(x,y)$ equal to ordinary squared error as in example 2 of section 3. It also equals Pearson's "mean-square contingency" and Light and Margolin's $R^2$, as they show in [18]. Goodman and Kruskal's $\lambda$ is $R^2$ for $S(x,y)$ equal to the prediction error (3.11).

All of the Goodman-Kruskal measures as well as Light and Margolin's are defined for multichotomous responses, not just the binary case. The theory developed in this paper extends easily to the multichotomous case, at the expense of more elaborate notation. It is only necessary to change the horizontal axis [0,1] in figure 1 into the usual barycentric simplex. Some details are given in Savage [21].

The interaction between the dichotomous and multichotomous cases is interesting in its own right. Suppose we have a trichotomy $\{1,2,3\}$ with probabilities $p_1,p_2,p_3$. This can be thought of as two successive dichotomies, first "{1} versus {2,3}" with probabilities $p_1$ and $p_{23} = p_2 + p_3$, and then in the latter case "{2} versus {3}" with probabilities $p_2/p_{23}$ and $p_3/p_{23}$. Starting with a measure $\sigma(p)$ of binary variation, this yields the measure $\tilde{\sigma}(p_1,p_2,p_3) = \sigma(p_1) + p_{23} \sigma(p_2/p_{23})$ of trinary variation. There are actually three such measures $\tilde{\sigma}(p_1,p_2,p_3)$, depending on which of the three categories $\{1,2,3\}$ is given the preferred
place at the beginning of the construction. The requirement that all three be identical is

\[ \sigma(p_1) + p_2\sigma(p_2/p_{23}) = \sigma(p_2) + p_{13}\sigma(p_1/p_{13}) = \sigma(p_3) + p_{12}\sigma(p_1/p_{12}). \tag{5.11} \]

**Theorem 2.** Constant multiples of the entropy \( \sigma_1(p) \) are the only internal variation functions (2.12) satisfying (5.11) for all choices of \( p_1, p_2, p_3 \).

**Proof.** If \( p_2 \) is varied while \( p_{23} \) is held fixed, then \( dp_1/dp_2 = -1 \), \( dp_1/dp_2 = 0 \). Differentiating the far left and far right sides of (5.11) gives

\[ \sigma'(p_2/p_{23}) = -\sigma'(p_3) - (p_1/p_{12})\sigma'(p_1/p_{12}) + \sigma'(p_1/p_{12}) \cdot (p_2/p_{23}). \tag{5.12} \]

Substituting from (4.12) gives

\[ s(p_1/p_{12}) = \sigma'(p_3) + \sigma'(p_1/p_{12}) + \sigma'(p_1/p_{23}) \cdot \sigma'(p_2/p_{23}) \cdot (p_2/p_{23}). \tag{5.13} \]

Now differentiate (5.13) with respect to \( p_3 \), holding \( p_1/p_{12} = c \) constant (so \( dp_1/dp_3 = -c, dp_2/dp_3 = c-1, \sigma'(p_2/p_{23})/dp_3 = -c/p_{23}^2 \)) giving

\[ 0 = \sigma''(p_3) - \sigma''(p_2/p_{23}) \cdot (c/p_{23}^2). \tag{5.14} \]

Taking \( p_2 = p_3 \) gives

\[ \sigma''(p_3) = c\sigma''(1/2)/p_{23}^2. \tag{5.15} \]
Integrating twice shows that \( \sigma(p_2) \approx [p_2 \log p_2 + (1-p_2) \log (1-p_2)] \).

6. **Comparison with the Ordinary ANOVA Situation.**

In the ordinary one-way ANOVA situation the components of the data vector \( \chi \), at (5.3), can take on any values, not just zero or one. Then the vector of group means \( \frac{\bar{x}}{p} \) and the vector of the grand mean \( \mu \) also are not confined to component values in \([0,1]\). Suppose we have a variation measure \( S(x, \mu) \), \( x \) and \( \mu \) any real numbers, which is used to define a measure of variation \( S(x, \mu) = \sum_{i,j} S(x_{ij}, \mu_{ij}) \) for vectors. When does the fundamental one-way ANOVA relationship (5.4) hold? The answer is that under reasonable conditions on the form of \( S(x, \mu) \), only constant multiples of squared error, \( c(x-\mu)^2 \), satisfy (5.4).

The major condition on the form of \( S(x, \mu) \) is analogous to (2.12). Letting \( \phi_m, A(x) \) represent the density of a \( \eta(m, A) \) random variable, define

\[
\sigma(m, A; \mu) = \int_{-\infty}^{\infty} S(x, \mu) \phi_m, A(x) dx. \tag{6.1}
\]

We require that

\[
\min_{\mu} \sigma(m, A; \mu) = \sigma(m, A; m). \tag{6.2}
\]

Roughly put, the obvious center \( m \) of a perfect \( \eta(m, A) \) histogram of data must also be the center in the sense of minimizing the total residual variation.
We also require that for each value of \( \mu \), \( S(x, \mu) \) be a non-decreasing function of \( |x - \mu| \), say

\[
S(x, \mu) = g_\mu(|x - \mu|), \quad \text{nondecreasing in } |x - \mu|, \quad (6.3)
\]

which is analogous to (2.4). Finally, for analytic reasons, we suppose there exists some positive constant \( C \) such that

\[
S(x, \mu) < e^{x^2/C} \quad (6.4)
\]

and

\[
\left| \frac{\partial S(x, \mu)}{\partial \mu} \right| < e^{x^2/C} \quad (6.5)
\]

for all \( \mu \) and almost all \( x \).

**Theorem 2.** Under conditions (6.2)-(6.5) the only variation functions

\[
S(x, \mu) = \sum_{i,j} S(x_{ij}, \mu_{ij})
\]

satisfying the one-way ANOVA relationship (5.4) are those with \( S(x, \mu) = c(x - \mu)^2 \).

**Proof.** Differentiating (6.1) with respect to \( m \) and setting \( \mu = m \) gives

\[
\left. \frac{\partial \sigma(m; A; \mu)}{\partial m} \right|_{\mu = m} = \int_{-\infty}^{\infty} g_m(|x - m|) \frac{x - m}{A} \Phi_m(x) dx = 0, \quad (6.6)
\]

the last equality by symmetry. Differentiation under the integral sign is justified for \( A < C \) by (6.4). (6.3) implies that \( \sigma(m; A; \mu) \) is differentiable in \( \mu \), and then by (6.2) we have
\[
\frac{\partial \sigma(m, A; \mu)}{\partial \mu} \bigg|_{\mu=m} = 0. \tag{6.7}
\]

Therefore for \( A < C \),

\[
\frac{\partial \tau(m, A; m)}{\partial m} = \frac{\partial \sigma(m, A; \mu)}{\partial m} \bigg|_{\mu=m} + \frac{\partial \sigma(m, A; \mu)}{\partial \mu} \bigg|_{\mu=m} = 0. \tag{6.8}
\]

The function \( \sigma(m, A; m) \) does not depend on \( m \), so we can make the definition

\[
\sigma(A) = \sigma(m, A; m) \tag{6.9}
\]

for \( A < C \).

Now substitute the following vector \( \underline{x} \) at (5.3). 1) All the \( n_i = J \), some integer. 2) The averages \( p_1, p_2, \ldots, p_I \) equal the expected values of order statistics of a random sample of size \( I \) from a \( \mathcal{N}(0, B) \) distribution. 3) The values \( x_{ij} - p_j, j = 1, 2, \ldots, J \), equal the expected values of the order statistics of a sample of size \( J \) from a \( \mathcal{N}(0, W) \) distribution. 4) \( B \) and \( W \) are any two positive numbers such that \( B + W < C \), the constant appearing in (6.4), (6.5).

Notice that the overall average \( p = 0 \) and that the histogram of all \( IJ \) values \( x_{ij} \) is approximately \( \mathcal{N}(0, B + W) \). Letting \( I \) and \( J \) get arbitrarily large it is possible to show, using (6.3) and (6.4), that

\[
\frac{1}{IJ} S(\underline{x}, \underline{p}) = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} e_0 (|x_{ij}|) \sigma(B + W). \tag{6.10}
\]
Likewise

$$\frac{1}{IJ} S(x, p_1) = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} g_{p_1}(|x_{i,j} - p_1|) \rightarrow \sigma(W). \quad (6.11)$$

and

$$\frac{1}{IJ} S(p_1, p_2) = \frac{J}{IJ} \sum_{i=1}^{I} g_{0}(|p_1|) \rightarrow \sigma(B). \quad (6.12)$$

Then (5.4) implies that

$$\sigma(B+W) = \sigma(B) + \sigma(W) \quad (6.13)$$

for $B+W < C$, the Cauchy-Hamel equation [1]. Since $\sigma(\cdot)$ is monotone increasing by (6.1), this implies that

$$\sigma(A) = cA$$

for some constant $c$. But for any choice of $m$, $\sigma(A) = \int_{0}^{\infty} S(x,m) \varphi_{m,A}(x)dx$, so completeness in $A$, with $m$ fixed, implies $S(x,m) = c(x-m)^2$ for almost all $x$. Since $S(x,m)$ is increasing in $|x-m|$ for $m$ fixed, actually $S(x,m) = c|x-m|^2$ for all $x$ and $m$. This completes the proof of theorem 3.

The contrast between theorems 1 and 3 show a basic difference between binary regression and regression with quantitative response. There are many $S(\cdot, \cdot)$ functions satisfying the one-way ANOVA table in the former case, but only squared error in the latter. In a similar vein it can be shown that the analogue of the reproduction property (4.20) for quantitative data holds only for squared error.
7. **More Complicated ANOVA Models.**

In table 2 we fitted an increasing series of linear logistic models by maximum likelihood and measured the residual variation after each fit by $S_1(\hat{\pi}, \pi) = -2 \log f_\pi(\hat{\pi})$. An important result, proved in a general exponential family framework by Gary Simon [23], shows that this process is additive in the same sense as the one-way ANOVA table. Let $\hat{\pi}^{(0)}, \hat{\pi}^{(1)}, \hat{\pi}^{(2)}, \ldots$ be the maximum likelihood estimates of $\pi$ under an increasing series of models

$$\log \frac{\pi_i}{X_i} = \beta_0 V_i^{(0)} + \beta_1 V_i^{(1)} + \beta_2 V_i^{(2)} + \ldots , \quad (7.1)$$

where $V_i^{(0)}, V_i^{(1)}, V_i^{(2)}, \ldots$ are an ordered sequence of covariate vectors.

(In the Toxoplasmosis example $V_i^{(0)} = 1, V_i^{(1)} = z_i, V_i^{(2)} = \pi_i, \ldots$.)

Then for any $a < b$

$$S_1(\hat{\pi}^{(b)}, \hat{\pi}^{(a)}) = S_1(\hat{\pi}^{(b)}, \pi^{(b-1)}) + S_1(\pi^{(b-1)}, \pi^{(b-2)}) + \ldots + S_1(\pi^{(a+1)}, \pi^{(a)}) . \quad (7.2)$$

None of the other measures of residual variation display this kind of additivity. On the other hand if the increasing models are linear spaces in the $n$-dimensional cube containing $\pi$, rather than linear spaces in the logits, then $S_2(\cdot, \cdot)$ rather than $S_1(\cdot, \cdot)$ adds correctly. (7.2) is an argument for using $S_1(\cdot, \cdot)$ in connection with linear logistic models, not necessarily in general.
Unfortunately the familiar orthogonality properties of ordinary ANOVA do not extend to the binary situation, even when using \( S_1(\cdot, \cdot) \) in combination with linear logistic models. Table 4 shows a two-factor

Penicillin Experiment (Eagle, Fleischman, Misselman, 1950)

<table>
<thead>
<tr>
<th>G</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>20/20*</td>
<td>20/20</td>
<td>19/20</td>
<td>17/20</td>
<td>19/20</td>
<td>15/20</td>
<td>11/20</td>
<td>4/20</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>20/20</td>
<td>19/20</td>
<td>18/20</td>
<td>19/20</td>
<td>17/20</td>
<td>11/20</td>
<td>9/20</td>
<td>6/20**</td>
<td>+1</td>
</tr>
</tbody>
</table>

\[ \log_2 \text{Dose in mg/kg} \]

\[ -3.5 \quad -2.5 \quad -1.5 \quad -1 \quad -0.5 \quad 0.5 \quad 1.5 \quad 2.5 \quad 3.5 \quad 4 \]

* actually 0/0  ** actually 3/10

Table 4. Penicillin experiment, from Table I of Eagle et al [7]. Mice were inoculated with varying doses of Pneumocci, then treated with a single injection of one of two Penicillin species. The data has been altered as indicated to make the two factors "Penicillin Species" and "Log Dose" orthogonal.

experiment which has been slightly doctored to give orthogonality. The two factors, penicillin species \( V^{(1)} \) and log dose Pneumococi \( V^{(2)} \), were coded as shown in Table 4. Table 5 displays minus twice the maximized log likelihood after fitting various linear logistic models.
Table 5. Linear logistic models fitted to the data of table 4. Minus twice the maximized log likelihood, $-2 \log f^*$, is shown for each model. "Groups" refers to fitting the observed survival proportion within each of the 16 cells of table 4. The figures in parentheses are the squared errors, $\sum_2 (\cdot, \cdot)$, which are roughly comparable to $-2 \log f^*$ as lemma 1 suggests.

Figure 3 illustrates the situation. The order in which $V^{(1)}$ and $V^{(2)}$ are introduced into the model does indeed make a difference, for example the difference between 1.874 and 1.242. The "local" theory described next says that this difference disappears asymptotically in large samples. The penicillin example is particularly non-local because $V^{(2)}$ has such a large effect. The usual experience of the author has been that orthogonal factors can be introduced in any order without significant changes in the ANOVA table. Nevertheless it is a disappointment that familiar orthogonal models such as the two-way layout don't work out perfectly.
Figure 3. An illustration of the results in table 5. It matters in which order \( V^{(1)} \) and \( V^{(2)} \) enter the model, even though they are orthogonal. For example the "Sum of Squares" for \( V^{(1)} \) is 1.242 if introduced first, but 1.874 if introduced second. Degrees of freedom are indicated in parentheses.

We now consider a particularly easy large sample situation which illustrated the special role of \( S_{11}(\cdot, \cdot) \). We suppose that the data are observed in groups of size \( n_1, n_2, \ldots, n_I \) respectively, the observed proportions of 1's being \( \mathbf{p} = (p_1, p_2, \ldots, p_I) \). (Notice the slight change of notation from (5.3.).) It is assumed that no covariate information is available within groups so of necessity the hypothesized models for the true probabilities assign constant values within groups,
\( \pi = (\pi_1, \pi_2, \ldots, \pi_I) \), where \( \pi \equiv \pi(\beta) \) depends in some smooth way on a lower dimensional parameter vector \( \tilde{\beta} \).

Keeping \( I \) fixed and letting the \( n_i \to \infty \) gives

\[ \tilde{\pi} = \pi(\tilde{\beta}) + \xi, \quad \xi \sim \mathcal{N}_I(0, \Sigma_{\pi_i}^{-1} \Sigma_{\pi_i} x_i / n_i), \quad (7.3) \]

\( \Sigma_{\pi_i} x_i / n_i \) indicating the diagonal matrix with \( i^{th} \) diagonal element \( \pi_i(\beta) x_i(\beta) / n_i \). If the true (but unknown) value of \( \beta \) is \( \beta^0 \) say, \( (7.3) \) gives an approximate linear model by \( I \) locally expanding \( \pi(\beta) \) as a linear Taylor series in \( \beta - \beta^0 \), and 2) Considering \( \xi \) to have the distribution connected with \( \beta^0 \).

(5.7) shows that

\[ \sum_{i=1}^{I} \sum_{j=1}^{J} S(x_{ij}, \pi_i(\tilde{\beta})) = \sum_{i=1}^{I} n_i S(p_i, \pi_i(\tilde{\beta})) + \sum_{i=1}^{I} n_1 \sigma(p_i). \quad (7.4) \]

It is clear that minimizing \( \sum_{i=1}^{I} \sum_{j=1}^{J} S(x_{ij}, \pi_i(\tilde{\beta})) \) over the choice of \( \tilde{\pi} \) is equivalent to minimizing \( \sum n_i S(p_i, \pi_i(\tilde{\beta})) \). The local theory depends on the following lemma.

**Lemma 3.** If \( \sigma(\cdot) \) is continuously twice differentiable then

\[ S(p, \pi) = -\sigma''(\pi) \frac{(p-\pi)^2}{2} + o(p-\pi)^2. \quad (7.5) \]

**Proof.** By Taylor's theorem

34
\[ S(p, \pi) = S(\pi, \pi) + \frac{\partial S(p, \pi)}{\partial p} \bigg|_{p=\pi} (p-\pi) + \frac{\partial^2 S(p, \pi)}{\partial p^2} \bigg|_{p=\pi} \frac{(p-\pi)^2}{2} + \cdots. \]

(7.6)

The lemma follows from (4.3), (4.8), and (4.11).

**Lemma 3** gives

\[ \sum_{n=1}^{\frac{1}{n}} S(p_i, \pi_i) \approx -\sum_{i=1}^{\frac{1}{n}} \sigma''(\pi_i^0)(p_i-\pi_i)^2 \]

(7.7)

for \( p \) and \( \pi(\beta) \) near \( \pi^0 = \pi(\beta^0) \). The ANOVA table based on minimized values of \( S(x, \pi(\beta)) \) is locally equivalent to that based on the quadratic form

\[ (p-\pi(\beta))D_{n^1_1} \sigma''(\pi_i^0)(p-\pi(\beta))'. \]

(7.8)

Such tables are locally additive and locally independent of the order of entering orthogonal covariates. (The local theory is appropriate to the case where \( n_i \to \infty \) because the fitted vectors \( \pi(\beta) \) as well as \( \pi \) are approaching \( \pi(\beta^0) \). "Orthogonal" here means with respect to the inner product \( D_{n^1_1} \sigma''(\pi_i^0)' \).

(7.3) implies that the locally efficient ANOVA tests and estimates are based on minimizing

\[ (p-\pi(\beta))D_{n^1_1/\pi_i^0}X_i(\pi_i-\pi(\beta))'. \]

(7.9)

Comparing (7.9) with (7.8) shows that the condition

\[ \sigma''(\pi) = \frac{c}{\pi X} \]

(7.10)
for some positive constant $c$ is necessary and sufficient for local
efficiency of ANOVA tables based on minimizing $S(x_i, \hat{g}(p))$. This condition
is equivalent to $\sigma(p) = c\sigma(p)$. Among the variation measures developed
in section 2, only constant multiples of the entropy give ANOVA tables
with the correct asymptotic properties.

Consider table 2 again. Comparing $SS_{CUBIC}$ to a $\chi^2_1$ distribution
is valid according to the asymptotic theory based on (7.3), (7.7) (with
$S(p, \pi) = S_1(p, \pi)$). This is just Wilks' maximum likelihood ratio criterion.
However comparing $MS_{CUBIC}/MS_{GROUPS}$ to an $F_{1, 30}$ distribution is also
valid. The difference is that the latter test is robust against a random
effects model in which there is no genuine cubic term, but where the true
probabilities for the cities are randomly scattered about a lower order
regression curve. This point is discussed in Pierce and Sands [19] and
also chapter four of Finney [9]. In the Toxoplasmosis example data from
the other age groups strongly suggests that the cubic effect indicated by
Wilks' test is spurious. See [20].

The last row of table 2 cannot be used to construct an approximate
$F$ test, at least not in the obvious way. The impediment is that
$SS_{\text{ERROR}} = S_1(x_i, \pi)$ (groups) is non-local in character. For the Toxoplasmosis
data it can be calculated that the expected value of $SS_{\text{ERROR}}$ is about
682 under the null hypothesis that all 697 subjects have the same prob-
ability of testing positive. This is 18% larger than the expected value
663 = 697 - 34 suggested by the local theory. Under the null hypothesis,
$F = MS_{CUBIC}/MS_{\text{ERROR}}$ will be smaller than an $F_{1, 663}$ variate. A correction
factor can be introduced but the author hasn't discovered any particularly
useful properties of the resulting test. Light and Margolin [17,18] suggest using a similar $F$ statistic, based on $S_2(\cdot,\cdot)$ rather than $S_1(\cdot,\cdot)$. However they only discuss the one-way ANOVA situation which, as we have seen, is rather special, so it is difficult to compare results.

In the Toxoplasmosis example it is reasonable to question how well the asymptotic theory applies to $SS_{\text{GROUPS}}$ since the numbers sampled are small in several of the groups. A numerical calculation shows that under the null hypothesis $H_0$: "the cubic regression model fits the true city probabilities perfectly", the expected value of $SS_{\text{GROUPS}}$ is about 34. This compares to 30 from the asymptotic theory. The implication is that $F_{\text{GROUPS}} = MS_{\text{CUBIC}}/MS_{\text{GROUPS}}$ may have a null distribution slightly smaller than $F_{1,30}$, but not seriously so particularly since the test accepts $H_0$ in this case.

The author is investigating the modified $F$ statistic

$$\frac{MS_{\text{of interest}}}{SS_{\text{GROUPS}}/E_{SS_{\text{GROUPS}}}}$$

(7.10)

$E_0$, indicating exception under the null hypothesis of no unexplained regression effects, which may be useful for grouped situations where the group sizes are small. A larger question raised in [18] is how well any of the usual asymptotic calculations work in small samples.
8. Estimation.

As in the previous section we suppose that the data are observed in
I groups of sizes \(n_1, n_2, \ldots, n_I\), \(\pi = (p_1, p_2, \ldots, p_I)\) and \(\pi = (\pi_1, \pi_2, \ldots, \pi_I)\) being the observed proportions of 1's and the true probabilities respectively. Some smooth model \(\pi = \pi(\beta)\) is assumed to hold, but \(\beta = (\beta_1, \ldots, \beta_K)\) will not always be indicated in the notation. In order to estimate \(\beta\) we consider minimizing a variation measure

\[
Q(\beta) = \frac{1}{2} \sum_{i=1}^{I} W(\pi_i, \beta) \left( \sum_{j=1}^{n_i} S(x_{1j}, \pi_i, \beta) \right)
\]

\[= \sum_{i=1}^{I} n_i W(\pi_i, \beta) \sigma(p_i, \pi_i, \beta), \tag{8.1}\]

where \(W(\cdot)\) is some positive differentiable function. The choice \(S(x, \pi) = S_1(x, \pi), W(\pi) = 1\) corresponds to maximum likelihood estimation of \(\beta\).

Setting derivatives equal to zero in (8.1) gives, after some manipulation involving (4.1)-(4.7), the estimation equations

\[
\frac{\partial Q(\beta)}{\partial \beta_k} = \sum_{i=1}^{I} n_i (p_i - \pi_i) A_i \frac{\partial \pi_i}{\partial \beta_k} + \sum_{i=1}^{I} n_i B_i \frac{\partial \pi_i}{\partial \beta_k} \quad k = 1, 2, \ldots, K, \tag{8.2}\]

where

\[
A_i = [W(\pi_i, \sigma(\pi_i))']', \quad B_i = W'(\pi_i, \sigma(\pi_i)). \tag{8.3}\]

The prime in (8.3) indicates differentiation, \(\sigma(\cdot)\) being assumed twice differentiable.
For a given value of $\beta$, say $\beta$, all of the vectors $\mathbf{p}$ satisfying (8.1) and therefore having $\beta$ as a solution to the estimating equations, lie in a linear space of dimension $n-K$, say $\mathbf{L}(\beta)$. This space is orthogonal to the $K$ vectors $D_{n_1A_1} \partial \pi / \partial \beta_k$, where $D_{n_1A_1}$ is the diagonal matrix with diagonal elements $n_1A_1$ and $\partial \pi / \partial \beta_k$ is the vector with elements $\partial \pi / \partial \beta_k$. See figure 4.

Figure 4. Estimation methods based on minimizing variation measures of form (8.1) give linear spaces of constant estimate. The one-parameter case is illustrated. In order for the method to be consistent and efficient, $\mathbf{L}(\beta)$ must pass through $\pi(\beta)$ and be perpendicular to the vector with $i$th coordinate $[n_1/n(\beta)X_i(\beta)] \partial \pi / \partial \beta_i$. In other words it must be equivalent to maximum likelihood estimation. The curved line indicates those $\mathbf{p}$ having minimum chi-square estimate equal to $\beta$. 

39
"Fisher consistency" requires that if $\hat{E} = \hat{\pi} (\hat{E})$ then $\hat{E}$ is a solution to the estimation equations. From (8.2) and (8.3) we see that this happens in general if and only if $W(\cdot) = 0$, that is if $W(\cdot)$ is a constant which we might as well set equal to one. In the case where the sample sizes $n_i$ all go to infinity with $I$ fixed, Fisher consistency is equivalent to ordinary consistency.

Asymptotic efficiency of the estimate obtained by minimizing (8.1) requires that $\mathcal{L}(\hat{E})$ be orthogonal to $\frac{\partial}{\partial \pi_1} \sum_i X_i \frac{\partial^2 W}{\partial \pi_1^2}$, $k = 1, \ldots, K$, as it is for the maximum likelihood estimate. If not the asymptotic efficiency is reduced by a factor dependent on the angle between $\mathcal{L}(\hat{E})$ and the plane of equal maximum likelihood estimates. These results, which go back to Fisher, are discussed in Efron [8]. Without going into detail, the asymptotics are quite convincing here and tend to give accurate small-sample measures of efficiency.

If $\pi(\hat{E})$ is a linear logistic family then the planes $\mathcal{L}(\hat{E})$ of equal maximum likelihood estimates are parallel. In this case the mapping from $\hat{p}$ to the maximum likelihood estimate $\hat{E}$ is equivalent to the reduction to minimal sufficient statistics.

We can write (8.1) as

$$Q(\bar{E}) = \sum_{i=1}^{I} n_i W(\pi_1(\bar{E})) \left\{ S(p_i, \pi_1(\bar{E})) + \sigma(p_i) \right\}.$$  \hspace{1cm} (8.4)

If $W(\cdot)$ is constant then minimizing $Q(\bar{E})$ is the same as minimizing $\sum_{i=1}^{I} n_i S(p_i, \pi_1(\bar{E}))$. If not, minimizing $\sum n_i W(\pi_1(\bar{E})) S(p_i, \pi_1(\bar{E}))$
results in non-linear surfaces of constant estimate. For example
\[ W(p_1) = \frac{1}{\pi_1} \chi_1, \quad S(p_1, \pi_1) = (p_1 - \pi_1)^2 \]
gives minimum chi-square estimates. Figure 4 shows that the curve of \( p \) values with this estimate equal \( \hat{\pi} \)
is tangent to the maximum likelihood plane \( \mathcal{H}(\hat{\pi}) \) at the point \( \pi(\hat{\pi}) \).
Such methods have full asymptotic efficiency in the usual sense, but
their "second order efficiency" is less than that for maximum likelihood,
see [8].

Applied problems usually include a mixture of estimation, hypothesis
testing, and descriptive statistics. The superiority of maximum likelihood
estimates argues for the use of the variation measure \( S_{\|}(\chi, \pi) \), if it is
necessary to restrict analysis to only one such measure.
REFERENCES


