PRACTICAL INference IN TWO-PHASE REgression

By

DAVID V. HINKLEY

TECHNICAL REPORT NO. 162
OCTOBER 19, 1970

THIS RESEARCH WAS SPONSORED BY THE ARMY RESEARCH OFFICE
OFFICE OF NAVAL RESEARCH, AND AIR FORCE OFFICE OF
SCIENTIFIC RESEARCH BY CONTRACT NO.
N00014-67-A-0112-0053 (NR-042-267)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
PRACTICAL INference IN TWO-PHASE REgression

by

DAVID V. HINKLEY

TECHNICAL REPORT NO. 162
OCTOBER 19, 1970

PREPARED UNDER CONTRACT N00014-67-A-0112-0053
NR-042-267
OFFICE OF NAVAL RESEARCH

Reproduction in Whole or in Part is Permitted for
any Purpose of the United States Government

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
PRACTICAL INFERENCE IN TWO-PHASE REGRESSION

by

David V. Hinkley

1. Introduction.

Recently several people have been interested in inference about the regression intercept $\gamma$ in the two-phase regression model

$$E(x_i) = \begin{cases} 
\theta + \beta_0 (u_i - \gamma) & (i=1, \ldots, \tau) \\
\theta + \beta_1 (u_i - \gamma) & (i=\tau+1, \ldots, T) 
\end{cases}$$

where $u_1 \leq \gamma < u_{\tau+1}$ and the $x_i$ are assumed to be independent normal with constant variance $\sigma^2$. The parameters $\theta$, $\beta_0$, $\beta_1$, $\gamma$, $\tau$ (and usually $\sigma^2$) are all unknown. Hudson (1966) and Hinkley (1969) have discussed procedures for computing the maximum likelihood estimates. Sylwester (1965) in unpublished work has proved the asymptotic normality of these estimates, and Hinkley (1969) derived an asymptotic distribution for the maximum likelihood estimate (MLE) $\hat{\gamma}$ which is a better moderate sample approximation.

The present paper is directed toward the application of the estimation procedure and associated inference procedures. The motivation is a set of data that will be examined where a major hypothesis is that $\beta_1 = 0$, and where inference about $\gamma$ and $\theta$ is of interest. Indeed, one would suspect this to be a fairly typical two-phase problem. In sections 2 and 3 respectively we outline the procedures for computing the MLE's and constructing confidence intervals, in each case for $\beta_1 = 0$ and $\beta_1$ unknown. (Note that $\beta_1 = 0$ covers all cases where $\beta_1$ is known by a
rotation of the model about \((\gamma, \theta)\). Section 4 covers the construction of joint confidence regions for \(\gamma\) and \(\theta\). For the numerical example in Section 5 we examine the accuracy of distributional approximations by Monte Carlo.

2. Estimation Procedures.

The estimation procedure for unknown \(\beta_1\) has been discussed both by Hudson (1966) and Hinkley (1969). Here we give a summary of their procedure and set up the corresponding procedure for the case \(\beta_1 = 0\).

2.1 Unknown \(\beta_1\)

Let \(\tilde{\beta}_{ot}, \tilde{\beta}_{1t}\) and \(\tilde{\gamma}_t\) denote the unconstrained least squares estimates of \(\beta_0, \beta_1\) and \(\gamma\) for the model (1.1) conditional on \(\tau = t\). Then the MLE \(\hat{\gamma}\) maximizes the piecewise continuous function \(Z(y)\) where

\[
Z(y) = \frac{C_t - D_t(\tilde{\gamma}_t + y) + E_t \tilde{\gamma}_t y}{(C_t - 2D_t y + E_t y^2)} \cdot \frac{(\tilde{\beta}_{ot} - \tilde{\beta}_{1t})^2}{C_{uu}, T}
\]

\((u_t \leq y \leq u_{t+1}; t=2, \ldots, t-2)\)

The coefficients \(C_t, D_t\) and \(E_t\) are defined by

\[
C_t = C_{uu,t} C_{*}, C_{uu,t} + \frac{tt^{*}}{T} (\bar{u}_t C_{*}, C_{uu,t} + \bar{u}^{*2}_t C_{uu,t}),
\]

\[
D_t = \frac{tt^{*}}{T} (\bar{u}_t C_{*}, C_{uu,t} + \bar{u}^{*}_t C_{uu,t})
\]
and

\[ E_t = \frac{tt^*}{T}(C_{uu,t} + C_{uu,t}^*), \]

where \( \tilde{u}_t, C_{uu,t} \) and \( \tilde{u}_t^*, C_{uu,t}^* \) are the mean and corrected sum of squares of \( (u_{t}, \ldots, u_{t}) \) and \( (u_{t+1}, \ldots, u_{T}) \) respectively and \( t + t^* = T \).

The simplest way to compute \( \hat{\gamma} \) is as follows, starting with \( t = 2 \),

(i) if \( u_t \leq \widetilde{\gamma}_t \leq u_{t+1} \) compute \( Z(\widetilde{\gamma}_t) \)

(ii) if not, compute \( Z(u_t) \) and \( Z(u_{t+1}) \).

Doing this for \( t \) up to \( T-2 \), the overall maximum occurs at \( \hat{\gamma} \).

The MLE's \( \hat{\beta}_o \) and \( \hat{\beta}_l \) satisfy

\[
\begin{bmatrix}
C_{uu,t} + \frac{tt^*}{T}(\tilde{u}_t - y)^2 - \frac{tt^*}{T}(\tilde{u}_t - y)(\tilde{u}_t^* - y) \\
- \frac{tt^*}{T}(\tilde{u}_t - y)(\tilde{u}_t^* - y) & C_{uu,t}^* + \frac{tt^*}{T}(\tilde{u}_t^* - y)^2
\end{bmatrix}
\begin{bmatrix}
\hat{\beta}_o \\
\hat{\beta}_l
\end{bmatrix}
\]

\[
= \begin{bmatrix}
C_{uu,t} + \frac{tt^*}{T}(\tilde{u}_t - \hat{\gamma}_t)(\tilde{u}_t - y) - \frac{tt^*}{T}(\tilde{u}_t^* - \hat{\gamma}_t)(\tilde{u}_t - y) \\
- \frac{tt^*}{T}(\tilde{u}_t - \hat{\gamma}_t)(\tilde{u}_t^* - y) & C_{uu,t}^* + \frac{tt^*}{T}(\tilde{u}_t^* - \hat{\gamma}_t)(\tilde{u}_t^* - y)
\end{bmatrix}
\begin{bmatrix}
\tilde{\beta}_{to} \\
\tilde{\beta}_{lt}
\end{bmatrix}
\]

with \( t = \hat{t} \) and \( y = \hat{\gamma}_3 \) note that \( u_{\hat{t}} \leq \hat{\gamma} < u_{\hat{t}+1} \). If \( \hat{\gamma} = \hat{\gamma}_s \),

then \( (\hat{\beta}_o, \hat{\beta}_l) = (\tilde{\beta}_{os}, \tilde{\beta}_{ls}) \).
The MLE of $\theta$ is

$$\hat{\theta} = \bar{x}_T - \frac{1}{T} \{ \hat{\beta}_0 \hat{\tau}(\bar{u}_T - \gamma) + \hat{\beta}_1 \hat{\tau}^*(\bar{u}_T^* - \gamma) \},$$

and the residual sum of squares about the fitted regression is

$$(2.2) \quad S = \sum_{i=1}^{T} (x_i - \bar{x}_T)^2 - \hat{\beta}_0 \hat{\tau} c_{uu, T} - Z(\hat{\gamma}),$$

where $\bar{x}_T = \frac{1}{T} \sum_{i=1}^{T} x_i$. The unbiased variance estimate is

$$\hat{\sigma}^2 = S/(T - 4).$$

For completeness we note that the asymptotic normal distribution of

$(\hat{\theta}, \hat{\beta}_0, \hat{\beta}_1, \hat{\gamma})$ has variances

$$\text{var}(\hat{\theta}) = \sigma^2 \left\{ \frac{\tau \beta_0^2}{\tau \beta_1^2 (\beta_0 - \beta_1)^2} + \frac{\beta_0^2 (\bar{u}_T^* - \gamma)^2}{c_{uu, \tau}} + \frac{\beta_1^2 (\bar{u}_T^* - \gamma)^2}{c_{uu, \tau}} \right\},$$

$$\text{var}(\hat{\beta}_0) = \frac{\sigma^2}{c_{uu, \tau}}, \quad \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{c_{uu, \tau}}$$

and

$$(2.3) \quad \text{var}(\hat{\gamma}) = \frac{\sigma^2}{(\beta_0 - \beta_1)^2} \left\{ \frac{1}{\tau} + \frac{1}{\tau^*} + \frac{(\bar{u}_T - \gamma)^2}{c_{uu, \tau}} + \frac{(\bar{u}_T^* - \gamma)^2}{c_{uu, \tau}} \right\}.$$

2.2 $\beta_1$ Known and Zero.

When $\beta_1 = 0$ the above results can be simplified. The residual sum of squares due to fitting (1.1) conditional on $\beta_1 = 0, \tau = t$
\( \gamma = y \) is

\[
(2.4) \quad S_{t}^{2}(y) = \sum_{i=1}^{T} (x_{i} - \bar{x})^{2} - \beta_{ot}^{2} \frac{[C_{t}^{i} - D_{t}^{i}(\tilde{\gamma}_{t}^{i} + y) + E_{t}^{i}\tilde{\gamma}_{t}^{i}y]^{2}}{C_{t}^{i} - 2D_{t}^{i}y + E_{t}^{i}y^{2}}
\]

where \( \tilde{\beta}_{ot} \) is as before and \( \tilde{\gamma}_{t}^{i} \) is the unconstrained least squares estimate of \( \gamma \) conditional on \( t = t \) and \( \beta_{1} = 0 \). The coefficients \( C_{t}^{i}, D_{t}^{i} \) and \( E_{t}^{i} \) are

\[
C_{t}^{i} = C_{uu,t} + \frac{tt^{*}u_{t}^{2}}{T} \quad D_{t}^{i} = \frac{tt^{*}u_{t}}{T} \quad \text{and} \quad E_{t}^{i} = \frac{tt^{*}}{T}.
\]

Now the MLE maximizes the piecewise continuous function \( Z'(y) \) defined as the final term in (2.4) for \( u_{t} \leq y \leq u_{t+1} \), \( t = 2, \ldots, T-2 \). The procedure for computing \( \hat{\gamma} \) is the same as that in section 2.1, using \( Z'(y) \) instead of \( Z(y) \). The other estimates are

\[
\hat{\beta}_{o} = \frac{[C_{t}^{i} - D_{t}^{i}(\tilde{\gamma}_{t}^{i} + y) + E_{t}^{i}\tilde{\gamma}_{t}^{i}y]}{(C_{t}^{i} - 2D_{t}^{i}y + E_{t}^{i}y^{2})} \tilde{\beta}_{ot}
\]

and

\[
\hat{\theta} = \bar{X}_{t} - \frac{\hat{\beta}_{o}}{T} (\bar{u}_{t} - y),
\]

with \( t = \hat{\tau} \), \( y = \hat{\gamma} \); if \( \hat{\tau} = \hat{\gamma}_{s} \), \( \hat{\beta}_{o} = \tilde{\beta}_{os} \) and \( \hat{\theta} = \frac{1}{T-s} \sum_{i=s+1}^{T} x_{i} \).

The residual sum of squares about the fitted regression is

\[
S' = \sum_{i=1}^{T} (x_{i} - \bar{x})^{2} - Z'(\hat{\gamma}),
\]
with $T-3$ degrees of freedom.

Now the variances of the asymptotic normal distribution of $(\hat{\theta}, \hat{\beta}_0, \hat{\gamma})$ are

$$\text{var}(\hat{\theta}) = \frac{\sigma^2}{\tau}, \quad \text{var}(\hat{\beta}_0) = \frac{\sigma^2}{\gamma_{uu, \tau}}$$

and

$$\text{var}(\hat{\gamma}) = \frac{\sigma^2}{\beta_0} \left( \frac{1}{\tau} + \frac{1}{\tau^*} + \frac{(\bar{u}_{\tau}-\gamma)^2}{\gamma_{uu, \tau}} \right)$$

(2.5)

2.3 Testing $\beta_1$.

It is worth remarking here that there are two possible tests of the hypothesis $\beta_1 = 0$. In both cases we obviously have to fit the general model (1.1). First we have the standard t-test using the estimated variance of $\hat{\beta}_1$. Second there is the likelihood ratio test with statistic

$$V = \frac{S'-S}{S} (T-4),$$

whose null distribution is asymptotically $F_{1, T-4}$. The latter test does not involve the nuisance parameter $\tau$, nor does it involve computation of $\hat{\beta}_1$, but it requires fitting the model with $\beta_1 = 0$. In many situations, of course, it might be worth fitting both models if we suspect that $\beta_1 = 0$.

3. Confidence Interval Procedures.

Suppose now that we want confidence intervals for $\gamma, \theta$ and possibly the slope parameter(s). There are two obstacles to using the asymptotic normal distributions directly. First, the variances depend
on unknown parameters, particularly the variance of \( \hat{\gamma} \) which depends heavily on \( \beta_1 - \beta_0 \). Second, and more difficult to handle, the asymptotic normal distributions may be poor approximations. This again applies particularly to \( \hat{\gamma} \); in moderate samples the distributions of \( \hat{\beta}_0, \hat{\beta}_1 \) and \( \hat{\theta} \) are closely approximated by their asymptotic normal forms. Fortunately the asymptotic chi-square distributions of the likelihood ratio statistics appear to be more robust small sample approximations. We therefore consider the construction of confidence intervals for \( \gamma \) using the appropriate likelihood ratio statistic. Further discussion of the distributional problems based on a Monte Carlo study will be taken up in section 5, with reference to the numerical example we analyse.

3.1 Unknown \( \beta_1 \).

Consider first the two-sided confidence interval for \( \gamma \). By the construction of \( Z(\gamma) \), see Hinkley (1969), the two-sided likelihood ratio test of \( H_0^*: \gamma = \gamma_0 \) with size 1-\( \delta \) is to accept \( H_0^* \) if

\[
\frac{Z(\hat{\gamma}) - Z(\gamma_0)}{S} \cdot (T-4) \leq F_{1,T-4}(1-\delta),
\]

where \( F_{1,T-4}(1-\delta) \) is the upper 1-\( \delta \) point of the \( F_{1,T-4} \) distribution and \( S \) is the residual sum of squares (2.2) used to estimate \( \sigma^2 \). (We have used the asymptotic \( \chi^2 \) distributions of \( Z(\hat{\gamma}) - Z(\gamma_0) \) and \( S \) as approximations, so that strictly the size 1-\( \delta \) is approximate). The 1-\( \delta \) confidence region for \( \gamma \) is the set of \( \gamma_0 \) satisfying (3.1).
Since \( Z(y) \) is not convex the region is not a single interval. In practice it seems reasonable to take as an approximate confidence region the smallest interval containing the exact confidence region, unless the likelihood has two very distinct humps. (This has some justification in the asymptotic convexity of \( Z(y) \)). So then the confidence interval for \( \gamma \) is \( (\gamma, \bar{\gamma}) \) with \( \gamma \) and \( \bar{\gamma} \) the smallest and largest roots of

\[
Z(y) = Z(\gamma) - \frac{S}{T-4} F_{1,T-4}(1-\delta).
\]

By the definition of \( Z(y) \) in (2.1) we see that (3.2) is a quadratic in \( y \) for \( u_t \leq y \leq u_{t+1} \). Direct substitution from (2.1) gives

\[
L_t y^2 - 2M_t y + N_t = 0,
\]

where

\[
L_t = [Z(\gamma) - \frac{S}{T-4} F_{1,T-4}(1-\delta)] E_t - (D_t - E_t \gamma_t)^2 \left( \frac{\beta_{ot} - \beta_{1t}}{C_{uu,T}} \right)^2,
\]

\[
M_t = [Z(\gamma) - \frac{S}{T-4} F_{1,T-4}(1-\delta)] D_t - (D_t - E_t \gamma_t) (C_t - D_t \gamma_t) \left( \frac{\beta_{ot} - \beta_{1t}}{C_{uu,T}} \right)^2,
\]

and

\[
N_t = [Z(\gamma) - \frac{S}{T-4} F_{1,T-4}(1-\delta)] C_t - (C_t - D_t \gamma_t)^2 \left( \frac{\beta_{ot} - \beta_{1t}}{C_{uu,T}} \right)^2.
\]
The procedure for finding \( \gamma \) is, then, to solve (3.3) for \( t = 2, 3, \ldots \) until at least one root lies in the corresponding interval \([u_t, u_{t+1}]\). Then the least root is \( \gamma \). If no such root exists for \( t \leq \hat{t} \) then we must set \( \gamma = u_2 \). The procedure for finding \( \gamma \) is to work backwards from \( t = T-2 \) until a root of (3.3) lies in the corresponding \([u_t, u_{t+1}]\), the largest such root being \( \gamma \). If no such root exists for \( t \geq \hat{t} \) then \( \gamma = u_{T-2} \). Note that the two terminal values of \( t \) form a confidence interval for \( \tau \).

To get a lower one-sided confidence limit for \( \gamma \) we note that the likelihood ratio test of \( H_0^*: \gamma = \gamma_0 \) with alternative \( H_1^*: \gamma < \gamma_0 \) is to reject \( H_0^* \) if

\[
\max_{\gamma \leq \gamma_0} Z(\gamma) - Z(\gamma_0) > \ell,
\]

say. Now for large samples the maximum over \( \gamma \leq \gamma_0 \) is close to or equal to \( Z(\gamma_0) \) when \( \gamma_0 < \hat{\gamma} \), and when \( \gamma_0 > \hat{\gamma} \) the maximum is \( Z(\gamma) \). Therefore an asymptotically equivalent test statistic is

\[
W = \begin{cases} 
Z(\gamma) - Z(\gamma_0) & \gamma_0 \geq \hat{\gamma} \\
0 & \gamma_0 < \hat{\gamma}
\end{cases}
\]

By symmetry we have asymptotically that

\[
\Pr(W > \ell|H_0^*) = 1/2 \Pr(Z(\gamma) - Z(\gamma_0) > \ell|H_0^*)
\]

so that the 1-\( \delta \) lower one-sided confidence limit for \( \gamma \) is the lower bound of the 1-2\( \delta \) per cent two-sided confidence interval. The one-
sided upper confidence limit is similarly derived.

To obtain confidence intervals for \( \theta, \beta_0 \) and \( \beta_1 \) in moderate samples, the normal approximations to distributions of the MLE's can be used with estimated standard errors. See, however, the cautionary remarks in Section 5. The alternative for \( \theta \), say, is to maximize the likelihood over \( \beta_0, \beta_1 \) and \( \gamma \) for several values of \( \theta \) and interpolate the two solutions of the equation in \( \theta \) corresponding to (3.3). From a practical point of view there seems to be too little to gain by using such a complicated procedure. Joint confidence regions for \( \gamma \) and \( \theta \) are discussed in Section 4.

3.2 \( \beta_1 \) Known and Zero

The exact confidence region for \( \gamma \) when \( \beta_1 = 0 \) is the set of values of \( \gamma \) satisfying

\[
\frac{Z'(\hat{\gamma}) - Z'(y)}{S^{'2}} (T-3) \leq F_{1,T-3}(1-\alpha),
\]

with approximate confidence level 1-\( \alpha \). Adopting the same interval procedure as in Section 3.1, we replace (3.3) by

\[ L_t' y^2 - 2M_t' y + N_t' = 0, \]

where

\[ L_t' = [Z'(\hat{\gamma}) - \frac{S'}{T-3} F_{1,T-3}(1-\alpha)]E_t' - \hat{\beta}_{ot}'(D_t'E_t'\hat{\gamma}_t')^2 \]

\[ M_t' = [Z'(\hat{\gamma}) - \frac{S'}{T-3} F_{1,T-3}(1-\alpha)]D_t' - \hat{\beta}_{ot}'(D_t'E_t'\hat{\gamma}_t')(C_t'D_t'\hat{\gamma}_t') \]

and

10
(4.1) \[ Z'(y, w) = T(\bar{x}_m - w)^2 + \frac{(\bar{\beta}_{ot} \bar{u}_t + \bar{u}_t - y) (\bar{x}_t - w))^2}{C_{u_t, t + t (\bar{u}_t - y)^2}} \]

\[ (u_t \leq y \leq u_{t+1}, t=2, \ldots, T-2) . \]

When \( y \) and \( w \) are the true values of \( \gamma \) and \( \theta \)

\[ Z'(\gamma) = Z'(y, w) \]

is distributed asymptotically as \( \sigma^2 \chi^2 \). Hence the 1-\( \delta \) confidence region for \( (\gamma, \theta) \) is the set of \( (y, w) \) satisfying

\[ (4.2) \quad Z'(y, w) \geq Z'(\gamma) - \frac{2S^t}{T-3} F_{2, T-3}(1-\delta) . \]

Now we can either solve (4.2) for several values of \( w \), or for several values of \( y \). In either case we get a discrete set of points on the boundary of the confidence region. For given \( w \) the solution of (4.2) proceeds exactly as in Section 3, but for given \( y \) (4.2) defines a single quadratic for all \( w \). Note also that solving (4.2) for a given set of \( y \) values gives the exact confidence region. It is easy to see from (4.1) and (4.2) that the boundary points for a given \( y \) are the real solutions of

\[ (4.3) \quad G_t'(y)w^2 - 2H_t'(y)w + K_t'(y) = 0 , \]

where

\[ G_t'(y) = T[C_{u_t, t + \frac{t(t+1)}{2} (\bar{u}_t - y)^2} , \]

12
\[ N_t' = z'(\hat{\gamma}) - \frac{S_t'}{T-2} \frac{F_{1,T-2}(\hat{\gamma}, 1-\beta)}{t_{T-2}(C_t, D_t')^2} \]

The derivation of \( \gamma \) and \( \hat{\gamma} \) is the same as before. Similar remarks to those in Section 3.1 apply to one-sided confidence limits for \( \gamma \).

As with \( \beta_1 \) unknown, the normal approximations to the distributions of \( \theta \) and \( \beta_0 \) are adequate for moderate sample inference about \( \theta \) and \( \beta_0 \).

4. Joint Confidence Region for \( \gamma \) and \( \theta \)

In some situations it may be that a joint confidence region for \( \gamma \) and \( \theta \) is required. One such region is a rectangle formed by two individual confidence intervals, but this has two defects: (a) there are infinitely many such regions for any confidence level, and (b) \( \hat{\theta} \) and \( \hat{\gamma} \) are correlated, so the region will be conservative. The alternative is to use likelihood regions, whose construction we discuss here.

4.1 \( \beta_1 \) Known and Zero

The residual sum of squares due to fitting (1.1) conditional on \( \beta_1 = 0, \tau = t, \gamma = \gamma \) and \( \theta = w \) is

\[ S_t'(y,w) = \sum_{i=1}^{T} (x_i - \bar{x})^2 - \frac{\sum_{i=1}^{T} (x_i - \bar{x})(u_i - \bar{y})}{\sum_{i=1}^{T} (u_i - \bar{y})^2} \]

so that corresponding to \( z'(\gamma) \) in Section 2.2 we have the regression sum of squares.
\[ H_t^*(y) = T\{c_{uu,t} (\tilde{x}_T - \tilde{\beta}_{ot} \frac{t}{T}(\tilde{u}_t - y)) - \frac{tt^* (\tilde{u}_t - y)^2 x_t^*}{T} \} \]

and

\[ \kappa_t^*(y) = \{c_{uu,t} + t(\tilde{u}_t - y)^2\} \{ Z'(\gamma) - \frac{g_{1}'}{T-\bar{y}} f_{2,T-3}(1) + tx_t^2 \} \]

\[-(\tilde{\beta}_{ot} c_{uu,t} + t(\tilde{u}_t - y) x_t)^2 . \]

The procedure is then to solve (4.3) for a grid of values between \( u_2 \) and \( u_{T-2} \), real solutions giving boundary points. The whole grid should be covered since disjoint regions may occur, a reflection of the non-convex likelihood. This procedure is applied in Section 5.

4.2 Unknown \( \beta_1 \)

When \( \beta_1 \) is unknown the regression sum of squares corresponding to \( Z'(y,w) \) can be written

\[ Z(y,w) = -T(\tilde{x}_T - w)^2 - \frac{g_1^2}{T} c_{uu,T} \]

\[ + \frac{(\tilde{\beta}_{ot} c_{uu,t} + t(\tilde{u}_t - y)(\tilde{x}_T - w))^2}{c_{uu,t} + t(\tilde{u}_t - y)^2} \]

\[ + \frac{(\tilde{\beta}_{ot} c_{uu,t} + t(\tilde{u}_t - y)(\tilde{x}_T - w))^2}{c_{uu,t} + t(\tilde{u}_t - y)^2} , \]

\((u_t \leq y \leq u_{t+1}; t=2, \ldots, T-2) . \)
The confidence region is now the set of \((y, w)\) satisfying

\[ (4.5) \quad Z(y, w) \geq Z(\hat{y}) - \frac{2S}{T-4} F_{2, T-4} (1-\alpha), \]

so the procedure of Section 4.1 applies with \(G'_t(y), H'_t(y)\) and \(K'_t(y)\) in (4.3) replaced by the corresponding coefficients derived from (4.4) and (4.5). (Note that solution of (4.5) for given \(w\) would require solution of a series of quartic equations in \(y\).)

5. A Numerical Example

To illustrate the procedures described above we analyse a small set of data obtained from replicated experimental determination of the relationship between blood factor VII production and warfarin concentration. The data formed part of a study by Pool and Borchgrevink (1964), and was originally analyzed by Rupert G. Miller who kindly provided it for this further analysis. The regression is linearized by logarithmic transformation of the warfarin concentration, which is then the independent variable \(u\). The dependent variable \(x\) is the average of replicated factor VII determinations. Table 1 gives the resulting fifteen data points to which we fit model (1.1)
<table>
<thead>
<tr>
<th>$u$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0.370483</td>
</tr>
<tr>
<td>2.52288</td>
<td>0.537970</td>
</tr>
<tr>
<td>3.0</td>
<td>0.607684</td>
</tr>
<tr>
<td>3.52288</td>
<td>0.723323</td>
</tr>
<tr>
<td>4.0</td>
<td>0.761856</td>
</tr>
<tr>
<td>4.52288</td>
<td>0.892063</td>
</tr>
<tr>
<td>5.0</td>
<td>0.956707</td>
</tr>
<tr>
<td>5.52288</td>
<td>0.940349</td>
</tr>
<tr>
<td>6.0</td>
<td>0.898609</td>
</tr>
<tr>
<td>6.52288</td>
<td>0.953850</td>
</tr>
<tr>
<td>7.0</td>
<td>0.990334</td>
</tr>
<tr>
<td>7.52288</td>
<td>0.890291</td>
</tr>
<tr>
<td>8.0</td>
<td>0.990779</td>
</tr>
<tr>
<td>8.52288</td>
<td>1.050865</td>
</tr>
<tr>
<td>9.0</td>
<td>0.982785</td>
</tr>
</tbody>
</table>

Table 1. Two-phase regression data

In this particular case a data plot shows that $\beta_1$ is expected to be zero. Also the difference in slopes is large compared to variation, which suggests that asymptotic distribution theory for $\hat{\gamma}$ may provide reasonable approximates. However, the sample size (fifteen) is small.

5.1 Regression Analysis

We fitted the two-phase model both with $\beta_1$ assumed zero and with $\beta_1$ unknown using the procedures outlined in Section 2. The resulting estimates are given in Table 2 together with their estimated standard errors based on the residual sum of squares estimates of variance $\hat{\sigma}^2$. 
Figure 1. Residual sums of squares plot for numerical example.
Application of both tests of $\beta_1$ mentioned in Section 2.3 leads to acceptance of $\beta_1 = 0$.

<table>
<thead>
<tr>
<th>case</th>
<th>$\hat{\gamma}$</th>
<th>ese($\hat{\gamma}$)</th>
<th>$\hat{\theta}$</th>
<th>ese($\hat{\theta}$)</th>
<th>$\hat{\beta}_\alpha$</th>
<th>ese($\hat{\beta}_\alpha$)</th>
<th>$\hat{\sigma}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1 = 0$</td>
<td>4.878</td>
<td>0.2179</td>
<td>0.9617</td>
<td>0.0147</td>
<td>0.1935</td>
<td>0.0211</td>
<td>0.00196</td>
</tr>
<tr>
<td>$\beta_1$ unknown</td>
<td>4.652</td>
<td>0.2439</td>
<td>0.9180</td>
<td>0.0144</td>
<td>0.1935</td>
<td>0.0211</td>
<td>0.00166</td>
</tr>
</tbody>
</table>

Table 2. Regression estimates and estimated standard errors (ese's)

In each case we also computed the .90, .95 and .99 confidence intervals for $\gamma$ using the likelihood method of Section 3, and these are given in Table 3. To supplement this, Figure 1 is a graph of the marginal log likelihoods of $\gamma$, expressed in terms of regression sums of squares, for $\beta_1 = 0$ and $\beta_1$ unknown. The regression sum of squares for a single-line regression is also plotted as a reference level. The effect of knowing $\beta_1$

<table>
<thead>
<tr>
<th>case</th>
<th>.90</th>
<th>.95</th>
<th>.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1 = 0$</td>
<td>(4.548, 5.293)</td>
<td>(4.450, 5.389)</td>
<td>(4.253, 5.655)</td>
</tr>
<tr>
<td>$\beta_1$ unknown</td>
<td>(4.160, 5.104)</td>
<td>(4.068, 5.202)</td>
<td>(3.641, 5.441)</td>
</tr>
</tbody>
</table>

Table 3. Likelihood confidence intervals for $\gamma$.  

17
is clearly defined: confidence intervals for $\gamma$ shrink to the left of $\hat{\gamma}$, as we see in Table 3. This is natural, because information about the slope of the second line makes classification of points from the first regression easier, whereas incorrect classification of points from the second regression does not disturb the assumed slope of the second line. The asymmetry of the likelihood conditional on $\beta_1 = 0$ also suggests a possible bias in $\hat{\gamma}$ when $\beta_1$ is assumed zero. This we examine further in section 5.2.

The corresponding confidence intervals for $\gamma$ using the approximating normal distribution for $\hat{\gamma}$ are given in Table 4; the variances were estimated using $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\tau}$, $\hat{\gamma}$ and $\hat{\sigma}^2$ in (2.3) and (2.5). Clearly any manipulation of these intervals to overcome the nuisance parameter $\hat{\beta}_0 - \hat{\beta}_1$ would give intervals worse than those in Table 3, as we should expect.

<table>
<thead>
<tr>
<th>case</th>
<th>0.90</th>
<th>0.95</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1 = 0$</td>
<td>(4.489, 5.267)</td>
<td>(4.403, 5.353)</td>
<td>(4.212, 5.544)</td>
</tr>
<tr>
<td>$\beta_1$ unknown</td>
<td>(4.214, 5.090)</td>
<td>(4.115, 5.189)</td>
<td>(3.894, 5.410)</td>
</tr>
</tbody>
</table>

Table 4. Confidence intervals for $\gamma$ using normal approximation to distribution of $\hat{\gamma}$.

The joint confidence region procedure outlined in Section 4.1 was applied to get .90, .95, and .99 confidence regions for $(\gamma, \theta)$ with $\beta_1$ assumed zero. The boundaries are plotted in Figure 2. Shifts from one smooth curve to another occurs as $\gamma$ goes through each $u_i$, that is as a data point goes from one regression to another.
The right-hand section of the region shows that $\theta$ can be located more precisely when we are reasonably sure that all points to the right of a given value of $\gamma$ have the common mean $\theta$. The dramatic shape of the region indicates that the smooth approximations got by solving (4.2) for given $w$ would be quite different; this approximation is easily done by eye in Figure 2.

5.2 Small-sample Distributions.

The properties of the inference procedures discussed and applied above are based on asymptotic theory, and may not be accurate approximations in samples as small as that in section 5.1. We therefore investigated the small-sample distributions empirically, using 500 generated samples from the two-phase regression model (l.1) with parameter values as estimated from the data in section 5.1 with $\beta_1$ assumed zero, given in the first row of Table 2. The sample size was kept at 15, with $u_1, \ldots, u_{15}$ as in Table 1. In each simulated sample the two-phase regression was fitted both with $\beta_1$ known and with $\beta_1$ unknown.

Table 5 gives the empirical means of $\hat{\beta}_0$, $\hat{\theta}$, $\hat{\gamma}$ and $\hat{\sigma}^2$, the empirical variances of $\hat{\beta}_0$, $\hat{\theta}$ and $\hat{\gamma}$ and the corresponding theoretical variances given by (2.3) and (2.5). There is a clear but small positive bias in $\hat{\gamma}$ when $\beta_1$ is known, and the variance of $\hat{\gamma}$ is considerably larger than that of the approximating normal distribution. Figure 3 shows the non-normality of $\hat{\gamma}$ by plotting the empirical distribution against normal quantiles for the case $\beta_1$ known. The
Figure 3. Distributions of $\hat{\gamma}$.

- Hinkley's approximation
- normal approximation
- empirical
solid curve represents Hinkley's (1969) approximation, which also performs badly, albeit a little better in the tails.

<table>
<thead>
<tr>
<th>Case</th>
<th>E(\hat{\gamma})</th>
<th>E(\hat{\beta}_o)</th>
<th>E(\hat{\theta})</th>
<th>E(\hat{\theta}^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>\beta_1 known</td>
<td>4.911</td>
<td>0.1932</td>
<td>0.9854</td>
<td>0.00194</td>
</tr>
<tr>
<td>\beta_1 unknown</td>
<td>4.898</td>
<td>0.1932</td>
<td>0.9932</td>
<td>0.00193</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case</th>
<th>empirical</th>
<th>theoretical</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>var(\hat{\gamma})</td>
<td>var(\hat{\beta}_o)</td>
</tr>
<tr>
<td>\beta_1 known</td>
<td>0.0582</td>
<td>0.000508</td>
</tr>
<tr>
<td>\beta_1 unknown</td>
<td>0.0908</td>
<td>0.000508</td>
</tr>
</tbody>
</table>

Table 5. Empirical means and variances of regression estimates.

The distribution of \( \hat{\beta}_o \) agrees quite well with the normal approximation, with a slightly inflated variance. The estimate \( \hat{\theta} \) also behaves well when \( \beta_1 \) is known, but the effect of removing the constraint \( \beta_1 = 0 \) is quite drastic, as can be seen from variance in Table 5. The approximating normal distribution for \( \hat{\theta} \) is apparently as suspect as that for \( \hat{\gamma} \) in such small samples. The variance estimate \( \hat{\theta}^2 \) is very well behaved.

The small sample size also influences the empirical distribution of \( Z(\hat{\gamma}) - Z(\gamma) \), giving longer tails than the approximating \( \chi^2_1 \) distribution.
But to be fair to the likelihood ratio inference, we also examined the distribution of \( (Z(\hat{\gamma}) - Z(\gamma))/\delta_u^2 \). Table 6 gives the empirical power of the two-sided .90, .95 and .99 likelihood ratio tests of \( \gamma \) (\( \beta_1 \) unknown) for five values of \( \gamma \) including the true value \( \gamma = 4.878 \). The empirical size of the tests agrees quite well with the theoretical size. Note the asymmetry of the power function, indicating a bias toward accepting high values of \( \gamma \). This matches the apparent but small bias in \( \hat{\gamma} \) when \( \beta_1 \) is unknown; see Table 5. The reason must be the asymmetry of the model \((\gamma < \frac{1}{2} T)\), whose influence disappears in larger samples. Finally, it should be emphasised that samples do not have to be much larger than considered here in order that asymptotic distributions be good approximations. For example, when \( T = 25, \gamma = 8.5 \) and \( \beta_1 - \beta_0 = \sigma \) (compared to \( T = 15, \gamma \approx 4.9 \) and \( \beta_1 - \beta_0 \approx 1.4 \sigma \) for our data), the empirical distributions of \( Z(\hat{\gamma}) - Z(\gamma) \) and \( \hat{\epsilon} \) are very close to the asymptotic distributions, however, the distribution of \( \hat{\gamma} \) is more reluctant to converge.

<table>
<thead>
<tr>
<th>( \gamma )</th>
<th>.10</th>
<th>.05</th>
<th>.01</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.478</td>
<td>0.494</td>
<td>0.368</td>
<td>0.162</td>
</tr>
<tr>
<td>4.678</td>
<td>.264</td>
<td>.168</td>
<td>.048</td>
</tr>
<tr>
<td>4.878*</td>
<td>.134*</td>
<td>.062*</td>
<td>.012*</td>
</tr>
<tr>
<td>5.078</td>
<td>.196</td>
<td>.116</td>
<td>.024</td>
</tr>
<tr>
<td>5.278</td>
<td>.422</td>
<td>.292</td>
<td>.068</td>
</tr>
</tbody>
</table>

Table 6. Empirical power of .90, .95, and .99 likelihood ratio tests of \( \gamma \); * empirical size.
6. Related Problems.

It is difficult to attempt an exhaustive list of objectives in two-phase regression, just as in many statistical procedures. But variants and extensions of the problem considered here can be easily handled. For example, one might have $n > 1$ sets of data, the model for the $j$th set being

$$
E(x_{ij}) = \theta_j + \beta_{0j}(u_{ij} - \gamma_j) \quad (i=1, \ldots, \tau_j)
$$

$$
E(x_{ij}) = \theta_j + \beta_{1j}(u_{ij} - \gamma_j) \quad (i=\tau_j+1, \ldots, T_j),
$$

with some hypothetical structure on the parameters. In particular one might wish to test $\gamma_j = \gamma$, or $\theta_j = \theta$, etc. Consider the test of equal $\gamma_j$ with the assumption $\beta_{1j} = 0$. Then denoting individual regression sums of squares conditional on $\gamma_j = \gamma$ by $Z_j'(y)$, the likelihood ratio statistic for testing $H_\gamma: \gamma_1 = \gamma_2 = \ldots = \gamma_n$ is

$$
L_\gamma = \frac{\sum_j Z_j'((\gamma_j)) - \max_y \sum_j Z_j'(y)}{\sum \sum (x_{ij} - \bar{x}_{i,j}, j)^2 - \sum_j Z_j'(\gamma_j)} \cdot \frac{\sum_j (T_j - 3)}{(n-1)}
$$

with obvious notation. Asymptotically $L_\gamma$ has the $F_{n-1, \sum_j (T_j - 3)}$ distribution. The test for a specific common intersection $\gamma_0$ replaces $\max_y \sum_j Z_j'(y)$ by $\sum_j Z_j'(\gamma_0)$, and $n-1$ by $n$ in $L_\gamma$. Construction of confidence intervals for the common intersection would be complicated unless $u_{ij} = u_i$ and $T_j = T$, in which case the denominator...
nator of $Z_j(y)$ is the same for all $j$, and the equation

$$\sum_j \{Z_j(\hat{\gamma}_j) - Z_j(y)\} = A$$

is a quadratic in $y$ for $u_j \leq y \leq u_{j+1}$ as in the case $n=1$. The procedure described in Section 3 is then directly appropriate.

Hypotheses involving $\beta_{0j}$ and $\theta_j$ lead to cumbersome expressions, but numerical calculations are easily performed. Several such problems, but for known $\tau_j$, have been considered by Sprent (1961).

Another type of problem that may arise is that of alternative models for the data. For example, an alternative to the two-phase model applied in Section 5 might be an exponential response curve $x = a - bp^u (0 < p < 1)$. The pragmatic approach to testing the alternatives is to fit both models and compute the difference in residual sums of squares $D$, say. Then using the parameter estimates in a small simulation of both models, the empirical distributions of $D$ under both models can be found and the data value of $D$ assigned to one of those distributions. This idea has been used by D. A. Williams (1970) to test between polynomial and exponential regression models.

Finally, we stress that the two-phase regression model discussed here is constrained to be continuous ($u_\tau \leq \gamma < u_{\tau+1}$), which would usually be the case. Without that constraint on $\gamma$, $\hat{\gamma}$ is still consistent, but $\hat{\tau}$ is not. In fact without the constraint the problem is a generalization of the mean-shift problem studied by Hinkley (1970).
References


**DOCUMEN T CONTRO L DATA - R&D**

<table>
<thead>
<tr>
<th>1. ORIGINATING ACTIVITY (Corporate author)</th>
<th>2a. REPORT SECURITY CLASSIFICATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department of Statistics</td>
<td></td>
</tr>
<tr>
<td>Stanford University</td>
<td></td>
</tr>
<tr>
<td>Stanford, California</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2b. GROUP</td>
</tr>
</tbody>
</table>

3. REPORT TITLE

**PRACTICAL INFERENCE IN TWO-PHASE REGRESSION**

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
   TECHNICAL REPORT

5. AUTHOR(S) (Last name, First name, Initial)
   HINKLEY, DAVID V.

6. REPORT DATE
   October 19, 1970

7a. TOTAL NO. OF PAGES
   26

7b. NO. OF REFs
   7

8a. CONTRACT OR GRANT NO.
   N00014-67-A-0112-0053

b. PROJECT NO.
   NR-042-267

c. d.

8b. ORIGINATOR'S REPORT NUMBER(S)
   TECHNICAL REPORT NO. 162

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. AVAILABILITY/LIMITATION NOTICES
    DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
    Office of Naval Research
    Logistics and Math. Science Branch
    Washington, D.C.

13. ABSTRACT
    Inference about the intersection in two-phase regression is discussed, with emphasis on application. A numerical example is given, and some empirical small-sample results are analysed.
Two-phase regression
Confidence regions

<table>
<thead>
<tr>
<th>KEY WORDS</th>
<th>LINK A</th>
<th>LINK B</th>
<th>LINK C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ROLE</td>
<td>WT</td>
<td>ROLE</td>
</tr>
</tbody>
</table>

**INSTRUCTIONS**

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (corporate author) issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR’S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers (either by the originator or by the sponsor), also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

   (1) "Qualified requesters may obtain copies of this report from DDC."

   (2) "Foreign announcement and dissemination of this report by DDC is not authorized."

   (3) "U.S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through"

   (4) "U.S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through"

   (5) "All distribution of this report is controlled. Qualified DDC users shall request through"

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring (paying for) the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, roles, and weights is optional.