A Comparison of Alternative Models for the Demand for Medical Care

Naihua Duan, Willard G. Manning, Jr.,
Carl N. Morris, Joseph P. Newhouse

Rand

HEALTH INSURANCE EXPERIMENT SERIES

R-2754-HHS

A Comparison of Alternative Models for the Demand for Medical Care

Naihua Duan, Willard G. Manning, Jr.,
Carl N. Morris, Joseph P. Newhouse

January 1982

Prepared under a grant from the
U.S. Department of Health and Human Services

Rand
SANTA MONICA, CA. 90406

The research reported herein was performed pursuant to Grant No. 016B80 from the U.S. Department of Health and Human Services

Library of Congress Cataloging in Publication Data
Main entry under title:

A Comparison of alternative models for the demand for medical care.

"R-2754-HHS."
Bibliography: p.
1. Insurance, Health--United States--Mathematical models. 2. Medical care--United States--Utilization--Mathematical models. I. Duan, Naihua.
[DNLM: 1. Health services research--United States. 2. Insurance, Health--United States. 3. Models, Theoretical. W 84.3 C737]
HG9396.C65 338.4'33621'0724 81-23538
ISBN 0-8330-0383-6 AACR2

Published by The Rand Corporation

PREFACE

This report evaluates alternative statistical models of the demand for medical care. The work was undertaken as a part of the Rand Health Insurance Study, a large-scale social experiment designed to investigate the effects of alternative health insurance plans on the utilization of health services and on health status. Four other Rand reports also deal with statistical problems in estimating the demand for medical services: J. P. Newhouse, C. E. Phelps, and M. S. Marquis, *On Having Your Cake and Eating It Too: Econometric Problems in Estimating the Demand for Health Services*, R-1149-1-NC, October 1979 (also published in the *Journal of Econometrics*, Vol. 13, 1980, pp. 365-390); W. G. Manning, J. P. Newhouse, and J. E. Ware, Jr., *The Status of Health in Demand Estimation: Beyond Excellent, Good, Fair, and Poor*, R-2696-HHS, December 1980 (also to be published in *Economic Aspects of Health*, edited by Victor R. Fuchs, University of Chicago Press, in press); and W. G. Manning, C. N. Morris, J. P. Newhouse, et al., *A Two-Part Model of the Demand for Medical Care: Preliminary Results from the Health Insurance Study*, R-2705-HHS, forthcoming (also in *Health, Economics, and Health Economics*, edited by J. van der Gaag and M. Perlman, North Holland Publishing Company, 1981).

The present report is a companion study to J. P. Newhouse, W. G. Manning, C. N. Morris, et al., *Some Interim Results from a Controlled Trial in Health Insurance*, R-2847-HHS, January 1982. That report describes the design of the experiment and presents the initial results from the model that was finally selected. The present report discusses the estimation problems, the alternative models considered, and the choice of a final model. A Rand paper by J. P. Newhouse and R. W. Archibald, *Overview of Health Insurance Study Publications* (P-6221, November 1978), lists other Health Insurance Study reports.

The present report should be of interest to persons studying the demand for medical services as well as to specialists in applied econometrics and statistics.

The research reported herein was performed pursuant to Grant No. 016B80 from the U.S. Department of Health and Human Services, Washington, D.C. The opinions and conclusions expressed herein are solely those of the authors, and should not be construed as representing the opinions or policies of any agency of the United States Government.

SUMMARY

In analyzing the demand for medical care, one wants to develop models that not only permit reliable inferences about behavior, but also yield reliable forecasts about future demand. In both cases, "reliable" means either minimum mean-squared error or consistent and efficient. Reliability is a prerequisite to informed decisions about alternative national health insurance packages, whether public or private. The stakes in these decisions are substantial. The nature of health insurance has a major effect on the size and character of the 10 percent of GNP spent on personal medical care services.

The distribution of annual medical expenditures by person has at least three characteristics that impede reliable inferences and predictions. First, about 20 percent of the population have no expenses for medical care during any given year. Second, the remaining 80 percent have positive expenses highly skewed to the right. Through much of their range, the positive expenditures are approximately lognormal. Third, the right tail of the distribution is longer than the lognormal distribution, because of the 10 percent of the population that have hospital utilizations.

In this report, we examine how several alternative estimation techniques perform on such data. The results indicate that complex models, which more accurately reflect the character of the distributions of medical services, perform better than simple ones. The models examined include analysis of variance (ANOVA) and analysis of covariance (ANOCOVA) on untransformed expenses; one-part models that use two-parameter Box-Cox transformations of expenses; two-part models that use separate equations to estimate the probability of positive expenses and the level of (log) nonzero expenses; and finally, a four-part model with separate equations to estimate the probability of positive expenses, the probability of inpatient expenses conditional on positive medical expenses, the (log) level of positive expenses for people with only ambulatory expenses, and the (log) level of positive expenses for those who have some hospital utilization.

The three simpler models can lead to less reliable results. The results of ANOVA and ANOCOVA on untransformed expenses are unbiased but imprecise (even with over a thousand observations) because the distribution of expenses is so highly skewed. Although the Box-Cox transformations increase the precision of the estimates substantially by reducing the effect of the skewness in the data, the predictions are statistically inconsistent because of the large number of non-spenders and the deviation from lognormality in the right tail. The two-part model corrects the difficulty associated with the probability of a zero expenditure but still produces inconsistent predictions as a result of the departure from lognormality in the right tail. The four-part model improves on the one- and two-part models by estimating the cases with inpatient expenses separately. The four-part model is more robust than ANOVA and ANOCOVA on untransformed expenditures.

Because the alternative models make different predictions of expenses for different health insurance plans, it is important to distinguish among them. A split-sample analysis leads us to reject the ANOVA and ANOCOVA on untransformed expenditures, and the one-part model. However, the analysis fails to distinguish between the two- and four-part models. Nevertheless, we believe that the four-part model is the better model because it more accurately reflects the complexity of the distribution of medical expenses. Thus, its estimates are consistent, whereas those of the one- and two-part models are inconsistent. The estimates of the four-part model are more precise and robust than those of ANOVA and ANOCOVA on untransformed data. As additional data become available, we expect the superiority of the four-part model to become more apparent.

The analysis reported here was performed on preliminary data from the Rand Health Insurance Study, a randomized social experiment designed to assess the effect of alternative health insurance plans on health services utilization and health status.

ACKNOWLEDGMENTS

We would like to thank William Lisowski and Daniel Relles for their statistical programming help. Emmett Keeler, Arleen Leibowitz, William Rogers, John Rolph, and Wynand van de Ven have provided us with useful comments. Careful reviews by Arthur Dempster and Robert Bell have helped markedly to improve the report. Professor Bradley Efron of Stanford University deserves our special thanks for his continuing advice to watch our distributional assumptions.

CONTENTS

FIGURES

TABLES

Chapter 1
INTRODUCTION

Americans have been debating the merits of cost sharing for medical services for decades. Central to that debate have been questions about the effect of cost sharing on the demand for medical services and on the level of health status. Some have argued that cost sharing does not affect demand, but is merely a tax on the sick; others have argued that cost sharing deters necessary care. Still others believe that cost sharing prevents abuse of the system for trivial problems. The debate has persisted because the data available to address these questions are meager and flawed by problems such as selection effects and lack of information about health status. To remedy these data problems, the federal government sponsored a social experiment or controlled trial in health insurance, the Rand Health Insurance Study (HIS), which started in 1974 and will conclude in 1982. Data from the earlier years of the study are now available to the HIS staff for analysis.

Once experimental data were available, the first analytic problem was to find reliable point estimates of the impact of cost sharing on the use of health services.[1] In particular, we want to forecast the cost of alternative insurance packages, whether public or private. Reliability is important because our forecasts have major policy implications. Some models lead to grossly imprecise results. Such results could lead us to conclude that cost sharing does not reduce use, because random error obscures the true response. If demand does not respond to cost sharing, the optimal insurance policy has little or no cost sharing; but such a conclusion does not necessarily hold if demand responds. Thus, failing to detect a true response could lead to excessively generous insurance.

---

[1] By reliable we mean either consistent and efficient or minimum mean squared error.

By contrast, other models are inconsistent and lead to an overestimate of response. Such estimators could cause us to reject a more generous insurance package because it appears more costly than it is. Because the budgets of large public programs, such as Medicaid, are in the tens of billions of dollars, more reliable estimates will have substantial payoff in terms of better decisions. Even large private plans, such as those covering General Motors employees, pay health insurance premiums in the hundreds of millions of dollars. With budgets and premiums of these magnitudes, unreliable results could lead to policy errors with major financial impact.

In modelling the use of medical services, we faced a tradeoff between precision and bias. Some models that we used early in the analysis produced results that were consistent but very imprecise. When we developed more precise models, they produced inconsistent predictions. We then developed more elaborate models that would eliminate inconsistency without a substantial loss of precision. However, these more elaborate models run the risk of overfitting the observed data. If we were fitting random error rather than true response, our forecasts for the other populations would be unreliable. Our analytical problem then is to make the appropriate choice among the competing models.

The source of the estimation problems is the distribution of medical expenses. At least three characteristics make this distribution difficult to model. First, about 20 percent of the population have no expenses for medical care during any given year. Second, the remaining 80 percent have positive expenses that are highly skewed. Through much of their range, the positive expenditures are approximately lognormally distributed. Third, the far-right tail of the distribution is too long even for a lognormal distribution, because about 10 percent of all cases have hospital utilizations. We have not found any simple parametric distribution to represent that 10 percent.

In this report, we examine how several alternative estimation techniques perform on our data. Not surprisingly, the better models

explicitly account for the characteristics of the distribution of medical services. The models we examine include analysis of variance (ANOVA) and analysis of covariance (ANOCOVA) with untransformed expenses as a dependent variable; the one-part model, which uses a two-parameter Box-Cox transformation of expenses; the two-part model, which uses two separate equations to estimate the probability of positive expenses and the level of (log) positive expenses; and a four-part model with separate equations to estimate the probability of positive expenses, the probability of inpatient expenses conditional on having positive medical expenses, the (log) level of positive expenses for those individuals with only ambulatory expenses, and the (log) level of positive expenses for persons who have positive inpatient utilization. More detailed specifications of the models are given in Chapter 3.

All the models except the four-part one have unsatisfactory aspects. The results of ANOVA and ANOCOVA on untransformed expenses are unbiased but imprecise (even with over a thousand observations) because expenses are so highly skewed. Although the Box-Cox transformation in the one-part model, by reducing the effect of the skewness in the data, substantially increases the precision of the estimates, the forecasts from this model are inconsistent because of the large number of zero expenses and the lack of lognormality in the right tail. Two-part models correct the inconsistency due to zero expenses but still produce inconsistent forecasts as a result of the departure from lognormality in the right tail. The one- and two-part models do estimate the median expense reasonably well, but they underestimate the impact of hospitalization. Such underestimation can have disastrous effects on forecasts of mean expenditure for medical services, because the top 10 percent of the distribution accounts for more than half of the total medical expenses.

In comparing the models, we have used two criteria: statistical consistency and minimum mean squared error. With fairly large sample sizes (e.g., on the order of several thousand), it is reasonable to expect that a model should produce statistically consistent estimates.

The consistency criterion motivated our development of the more refined transformed models. On the other hand, with fixed (finite) sample sizes, there is a tradeoff between bias and variance to be considered. That is, a model that produces results that are slightly biased (or inconsistent) but more precise (having smaller variance) might be preferable to one free from bias at the expense of lower precision. We therefore also consider the mean squared error as a compromise between bias and variance. For this purpose, we have used a split-sample analysis to evaluate the various models in terms of mean squared (forecast) error. The split-sample analysis based on mean squared forecast error, and also on mean forecast bias, leads us to reject the ANOVA and ANOCOVA on untransformed expenditure, as well as the one-part model, but it fails to distinguish between the two- and four-part models. Nevertheless, we believe that the four-part model is the better one for analyzing our data because its estimates are consistent and more data will become available. By contrast, the two-part model is inconsistent, because departures from normality when using it are systematically related to the covariates.

* * *

Chapter 2 briefly describes the design of the Health Insurance Study, which is the source of our data, and the sample. Chapter 3 provides a rationale for and description of each of the models considered, and Chapter 4 indicates the sensitivity of the empirical results to the estimation model. Chapter 5 compares the models empirically in terms of forecast bias and mean squared forecast error, using a split-sample technique. Chapter 6 summarizes the findings of this study.

Chapter 2
THE DESIGN OF THE HEALTH INSURANCE STUDY, THE SAMPLE, AND THE DATA

The Health Insurance Study (HIS) is a social experiment designed to study how different health insurance policies affect the demand for health services and thus the health status of individuals. Among other goals, it seeks to determine the responsiveness of demand to varying degrees of cost sharing. Past studies of this subject have typically used nonexperimental data that suffer from several flaws: insurance is potentially endogenous; existing policies are difficult to describe parametrically; utilization data are frequently based on recall and are subject to reporting biases; and coinsurance rates and deductibles often vary little for a given service, such as hospitalization.[1] The HIS was designed to avoid these problems.

THE DESIGN

Because our intent in this report is to find reliable estimates of the effect of health insurance on demand, we will limit the description of the HIS design to the experimental insurance plans.[2] The families participating in the experiment were assigned to 14 different insurance plans.[3] About one-third of the sample were assigned to a plan with a zero coinsurance rate (they received free care). Nearly one-fifth faced a 25-percent coinsurance rate, subject to an upper limit on annual out-of-pocket family expenditures

---

[1] Newhouse (1978, 1981) provides a critical review of past studies and their methodological problems.

[2] Newhouse (1974) and Brook et al. (1979) provide fuller descriptions of the design. Newhouse et al. (1979) discuss the measurement issues for the second generation of social experiments, to which the HIS belongs.

[3] In addition, the HIS has two groups enrolled in a prepaid group practice or health maintenance organization. Results for these two groups will be discussed in subsequent reports.

of 5, 10, or 15 percent of the previous year's income, or $1000, whichever was less.[4] This limit was called the Maximum Dollar Expenditure (MDE). For some of this group, the coinsurance rate was 50 percent for dental and outpatient mental health services. Just under one-twelfth of the sampled families were enrolled in a plan having a 50-percent coinsurance rate for all services, subject to the MDE (in Dayton one-fifth of the sample has this plan). One-fifth of the sample were assigned to a plan with a 95-percent coinsurance rate, subject to the MDE. In effect, this last group of families had an income-related family deductible. Finally, about one-fifth of the families faced a 95-percent coinsurance rate for outpatient services, subject to $150 limit on out-of-pocket expenses per person ($450 per family). In this plan, all inpatient services were free, so that, in effect, this plan had an individual outpatient deductible.[5]

All plans covered a wide variety of services. The only significant exclusions were outpatient mental health services in excess of 52 visits per year, nonpreventive orthodontia, and cosmetic surgery unrelated to accidents occurring after the start of the experiment.[6] Dental services and outpatient mental health services were, however, treated somewhat differently in the first year in Dayton, and so analysis of those services will not be considered here.[7] The same coinsurance rate applied to all medical services, with the exceptions noted above.

---

[4]The limit was $750 for the 25-percent coinsurance plans in the Massachusetts and South Carolina sites.

[5]The coinsurance rate for the family and individual deductible plans was 100 percent in Dayton Year 1. The rate was changed to 95 percent to increase the incentive to file in all other site-years, although there was no statistical evidence of underfiling.

[6]In the case of each exclusion, it is questionable whether anything could have been learned about steady-state demand during the 3- to 5-year lifetime of the experiment.

[7]Dental services for adults were only covered in the plan with a zero coinsurance rate; expenditures on outpatient mental health services did not count toward the MDE. After Year 1 in Dayton and

The families were assigned to plans using the Finite Selection Model (Morris, 1979). This model is designed to achieve as much balance across plans as possible while retaining randomization; i.e., it makes the experimental plans orthogonal to the demographic covariates. The expected gain in precision from using this model rather than simple random assignment in this experiment is about 25 percent (Morris, 1979). (Random refusals of the enrollment offer, which were about 15 percent in this experiment, degrade the 25-percent gain in proportion to the refusal rate (Morris, Newhouse, and Archibald, 1979).) We have found no unintended differences between the enrolled group and the Dayton population, the only site for which this analysis is complete.

The sample is a random sample of each site's population, but the following groups were not eligible: (1) those 62 years of age and older; (2) those with incomes in excess of $25,000 (in 1973 dollars); (3) those eligible for the disability Medicare program; (4) those in jails; (5) those in the military or their dependents; and (6) those with service-connected disabilities.

THE SAMPLE

The sample used in this report consists of those enrollees who participated for a full year in the first 3 years of Dayton and for 2 years in Seattle and Massachusetts, plus those who died during any year. Excluded are other individuals with partial years of participation: newborns, adoptees, suspended participants (e.g., those who joined the military), participants who voluntarily attrited, and individuals who were involuntarily terminated for noncompliance during the year.[8] But a person who, say, attrited in Year 2 was included

---

all other sites, dental services for adults and outpatient mental health services (up to 52 visits per person annually) were covered like any other service in all plans.

[8]We expected these cases to behave differently from the full-year population. With the data available, there are not enough cases to analyze the differences with precision. Inclusion of an indicator variable for the condition would essentially dummy out the cases. Instead, we delay their analysis until we have enough data

in Year 1 if he participated for all of that year. Moreover, the inpatient expenditures on newborns were entirely allocated to the mother for the purposes of this analysis. The exclusions account for about 5 percent of the total sample. Table 2.1 contains the number of observations per site year. The decrease from year to year reflects death, attrition, termination, or suspension, net of births and adoptions.

THE DATA

Dependent Variables

We have confined our analysis to covered medical expenditures other than dental or outpatient mental health care. Thus, total medical expenses (MED) include all inpatient care, medical outpatient care, including services provided by nonphysicians such as chiropractors and optometrists, and (largely prescription) drugs and supplies. Claims filed by participants, including those for unreimbursed expenses, provide data on the amount and type of expenses. For some analysis we do distinguish inpatient users from ambulatory users. A user of inpatient services is anyone who has been admitted to or born in a hospital at any time during the year.

Insurance Plan Variables

We have employed an analysis of covariance (ANOCOVA) specification of the insurance plans, with four dummy variables, one each for a family medical coinsurance rate of 25 percent (P25); a family medical coinsurance rate of 50 percent (P50); a family medical coinsurance rate of 95 percent (behaviorally, approximately a family

---

to analyze them properly. We also excluded four cases who participated in the disability Medicare program. They were enrolled because the HIS expected to receive a waiver from the Social Security Administration (SSA) permitting participation. When SSA denied the waiver, individuals eligible for such benefits became ineligible for the HIS. The experiment enrolled no more such cases. We have excluded these four cases in order to maintain a defined population to which we can generalize.

Table 2.1

SAMPLE BY SITE-YEAR[a]

Site-Year	*Number in Sample*
Dayton 1	1110
Dayton 2	1103
Dayton 3	1092
Seattle 1[b]	1171
Seattle 2[b]	1146
Fitchburg 1	704
Fitchburg 2	693
Franklin 1	875
Franklin 2	871

[a]Full-year participants plus deaths; excludes newborns, adoptees, attritions, terminations, suspensions, and four SSI recipients (who were enrolled in anticipation of a waiver that was never granted).

[b]The Seattle sample includes only those individuals in the fee-for-service experiment. Results on participants enrolled in a health maintenance organization will be reported separately.

deductible) (PFD);[9] and the individual deductible of $150 per person or $450 per family for outpatient care (IDP). The free care plan is the omitted group.

In the four-part model, we have noticed a rather large interaction between age and plan. Adults respond to plan in their (conditional) probability of utilizing inpatient services, whereas children do not. Therefore, we have an interaction specification in that equation, with dummy variables for an adult on the 25-percent plan (AP25), an adult on the 50-percent plan (AP50), an adult on the family deductible plan (APFD), and an adult on the individual deductible plan (AIDP).

[9]A 100-percent coinsurance rate would be exactly a family deductible.

Other Covariates

The model specification includes covariates for other experimental treatments, age, sex, race, family income and size, prior contact with the medical system, self-reported health, pain, and worry. These and other covariates are given in Table 2.2 and are described more fully in Manning et al. (1981).

Except the interactions noted above or in Table 2.2, we have used no interactions. Although we have not tested for all possible interactions, we did examine some that are important for policy purposes (e.g., between income and plan) and found them statistically insignificant. In subsequent analysis we expect to test the linearity of the models using Pregibon's (1980) goodness-of-link test.

Table 2.2

INDEPENDENT VARIABLES

TREATMENT VARIABLES - INDICATOR (0,1) VARIABLES

P00 = 1	if free plan
P25 = 1	if medical coinsurance rate = 25 percent
P50 = 1	if medical coinsurance rate = 50 percent
PFD = 1	if medical coinsurance rate = 95 percent
IDP = 1	if individual deductible plan
EXAM = 1	if received physical exam at enrollment[a]
YR3 = 1	if enrolled for 3 years[b]
WEEKLY = 1	if filed health diary weekly[c]
NOHR = 1	if did not file health diary[c]

SOCIODEMOGRAPHIC VARIABLES

Indicator (0,1) Variables

BLACK = 1	if the family is black
FEMALE = 1	if female
CHILD = 1	if age < 18
FCHILD =	FEMALE · CHILD
AFDC = 1	if someone in family received Aid to Families with Dependent Children
NOMD = 1	if no regular physician for any family member
NOMDVIS = 1	if no visits to physician in past year
HLTHG = 1	if self-rated health is good[d]
HLTHFP = 1	if self-rated health is fair or poor[d]
PAINGS = 1	if in great pain or in some pain[e]
PAINL = 1	if in little pain[e]
WORRGS = 1	if health is of great worry or of some worry[f]
WORRL = 1	if health is of little worry[f]
NEWMEM = 1	if added to family after pre-entollment interviews[g]
FAD = 1	if female adult

Continuous Variables

LINC =	ln (average family income in 1972 dollars)[h]
LFAM =	ln (family size)
INMDVIS =	[max (1, number of baseline year physician visits)]$^{-1}$
MAGE, FAGE =	a function relating outpatient utilization to age and sex (MAGE for males and FAGE for females), based on National Center for Health Statistics data on physician visits

[a] No exam is the omitted category.

[b] Five years is the omitted category.

[c] Filing biweekly is the omitted category.

[d] Excellent health is the omitted category.

[e] No pain is the omitted category.

[f] No worry about health is the omitted category.

[g] All other unknown individual (not family) variables set to zero.

[h] Income is set equal to $1000, if reported to be less. The years averaged are 1972, 1973 in Dayton, and 1973, 1974 in Seattle, Fitchburg, and Franklin.

Chapter 3
THE ALTERNATIVE MODELS

Each of the models was developed as an alternative to correct the shortcomings in the preceding ones. Included among the shortcomings are a lack of precision in the untransformed ANOVA and ANOCOVA models and bias in the predictions made by the transformed models. The one-part model is an attempt to correct the precision problems with direct analysis of untransformed medical expenditures. The two-part model is an attempt to correct the inconsistency due to nonspenders in the one-part model. The four-part model is an attempt to correct the two-part model's inconsistency due to the characteristics of inpatient utilization. Thus, we expect the four-part model to be better than the two-part model, which, in turn, should be better than the one-part model. All three *should* be more precise than the direct analysis of untransformed expenses. Later, in Chapter 5, we will provide a formal evaluation of the models in terms of mean squared forecast error and mean forecast bias.

Let us examine each of these models, their rationales, and their liabilities.

ANALYSIS OF VARIANCE (ANOVA) ON UNTRANSFORMED EXPENSES

The simplest model for expenses by plan is the ANOVA model, with plans entered as indicators:

$$Y_i \equiv MED_i = \mu + \alpha_i + \varepsilon_i , \qquad i = 1, \ldots, N , \qquad (3.1)$$

where μ is the grand mean, α is the plan effect, and ε is the error. The sample mean for each plan is the estimate of the mean of the expenses on that plan.

ANOVA has the advantage that it yields unbiased forecasts. Whether the error term ε is normally distributed or not, the sample average is an unbiased estimate of the plan mean as long as the plan

assignment is independent of the error ε_i, which the design of the experiment attempts to ensure. In contrast, the transformed analyses discussed later yield biased and statistically inconsistent forecasts if the error distribution is misspecified.

On the other hand, if the error is not normally distributed, the ANOVA estimate can be very sensitive to extreme values, i.e., to large expenses. (Medical expenses are obviously bounded from below.) The distribution of the HIS medical expenses is highly skewed toward the positive side, as indicated by Fig. 3.1, a normal plot[1] of medical expenses for the free plan from Dayton Year 1. Because of the skewness, the sample average does not provide an efficient estimate of the plan mean.

Clearly, we cannot estimate the effects of covariates on medical expenses by using the ANOVA model, but there are many reasons to do so. Accounting for the effects of relevant covariates can improve the precision of the estimated plan effects, and can also remove possible (presumably small) bias due to any imbalance in the plan assignment. Moreover, the effects of certain covariates (e.g., income) are relevant to policy (e.g., distributional questions).

ANALYSIS OF COVARIANCE (ANOCOVA) ON UNTRANSFORMED EXPENDITURES

The ANOVA model can be improved by including covariates known to affect medical expenses. A plausible model is the analysis of covariance (ANOCOVA) model

$$Y_i = x_i\beta_1 + \varepsilon_i \ , \qquad (3.2)$$

[1]The normal plot is sometimes called a Q-Q (Q for quantiles) plot. These figures plot the quantiles $F^{-1}(p)$ of the empirical distribution against the quantiles of a normal distribution with the same mean and variance. If the empirical distribution is normal, the quantiles will have the same values, and a plot of the quantiles will fall on a 45-degree line. In these plots, the axes are measured in 1σ units as deviations from the mean.

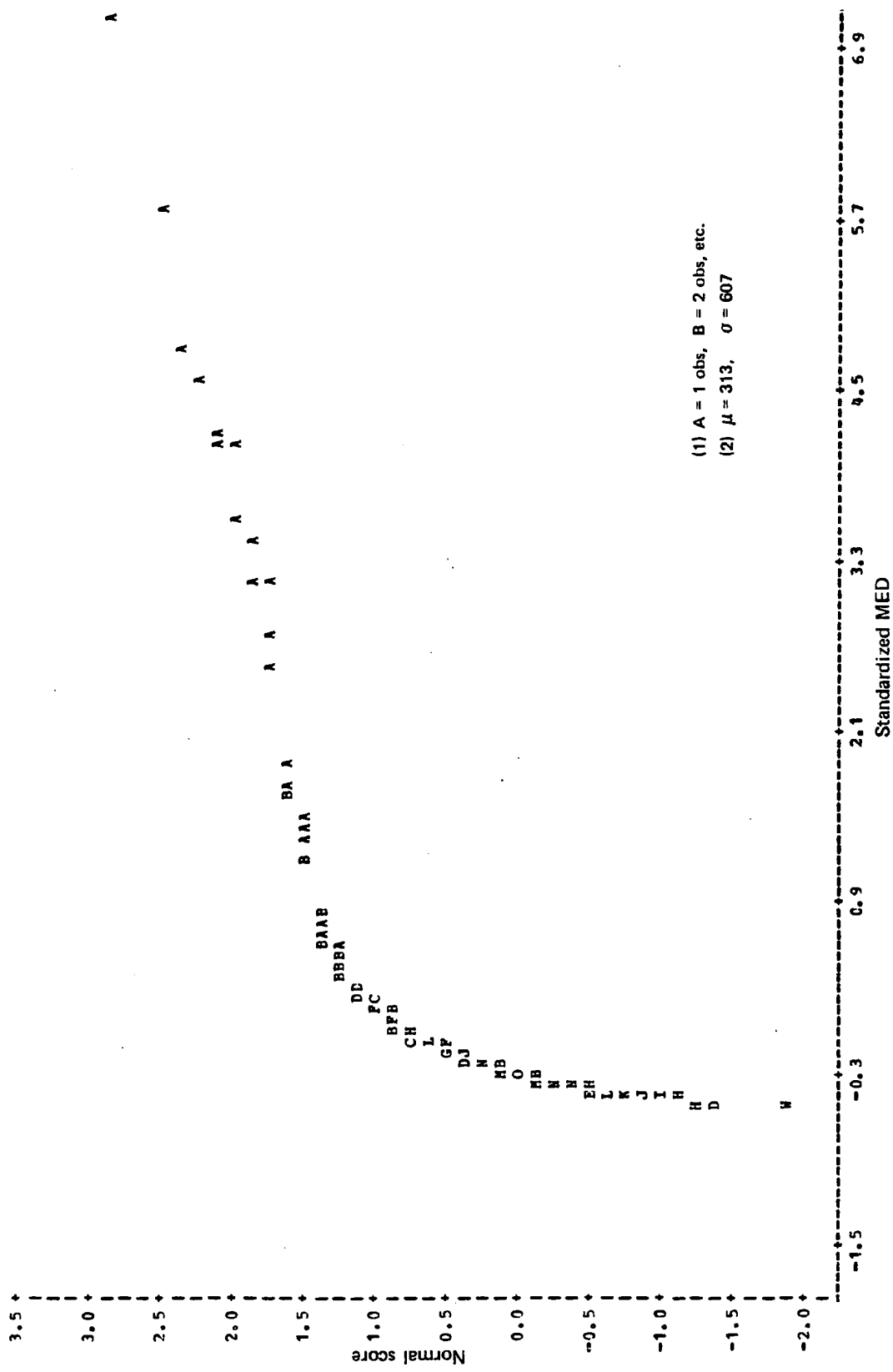


Fig. 3.1--Normal plot of the medical expenses for the free plan, Dayton Year 1

where x_i is a row vector of explanatory variables, including plan indicators and other covariates, and β_1 is a column vector of coefficients to be estimated. The ordinary least squares (OLS) estimate of the regression coefficients β_1 is

$$\hat{\beta}_1 = (X'X)^{-1}X'Y \ .$$

ANOCOVA has the advantage of yielding unbiased forecasts if the true model is linear [E (Y) = Xβ], and if the error term is independent of X. The disadvantage with ANOCOVA is that, like ANOVA, it is sensitive to extreme values. The distribution of the residuals after fitting the ANOCOVA model is highly skewed, as indicated by Fig. 3.2, a normal plot of the residuals for Dayton Year 1. Therefore the ANOCOVA estimate is not efficient.

ONE-PART MODEL

In the third model, we take a logarithmic transformation of medical expenses to diminish the influence of the extreme values, and analyze the linear model on the log scale:

$$\log (MED_i + \$5) = x_i\beta_3 + \varepsilon_{3i} \ . \qquad (3.3a)$$

After taking the logarithmic transformation, most of the distribution, including the right tail, looks roughly normal. A constant, namely \$5, is added so that we are not taking the log of \$0 for the nonspenders. The value \$5 is chosen because it minimizes the skewness of residuals. This transformation is (nearly) the best in the family of power transformations:

$$f(y) = (y + c)^p \qquad \text{if } p \neq 0 \ ,$$

$$f(y) = \ln (y + c) \qquad \text{if } p = 0 \ .$$

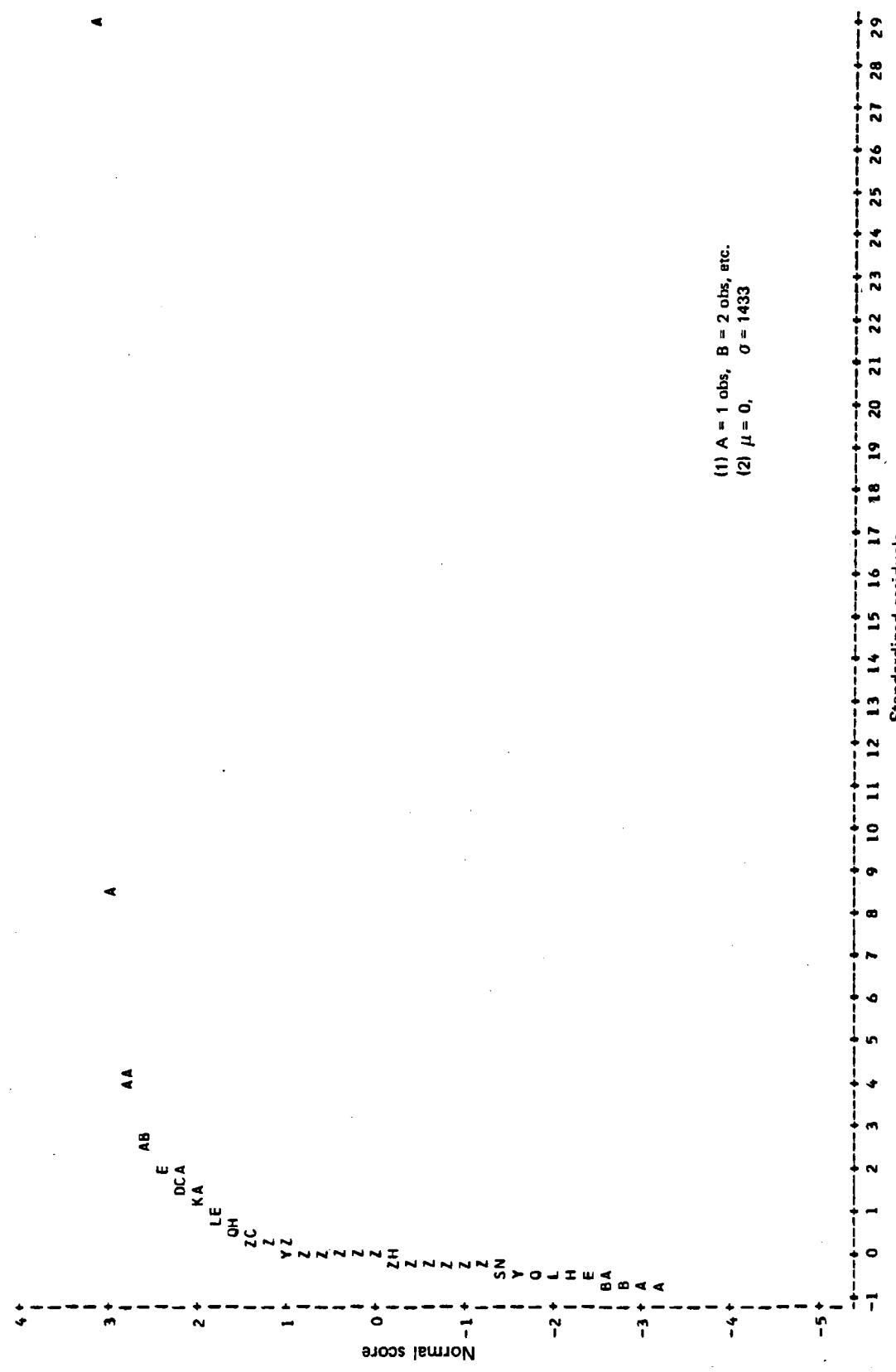


Fig. 3.2--Normal plot of residuals from ANOCOVA on untransformed expenses, Dayton Year 1

The objective here is to make the dependent variable $y = f(y)$ close to normal; see Box and Cox (1962). Figure 3.3 is a normal plot of the transformed dependent variable ln (MED + 5) for the free plan in the first year of Dayton. The distribution of the transformed variable is much closer to being normal than the raw expenditures in Fig. 3.1.[2] Reducing the departures from normality makes the estimates more robust. In addition, the estimated coefficients will be more precise with the log transform because the coefficient of variation is decreased under the assumption of log normality.[3]

Under this model, the expected medical expense for an individual with characteristics x_i is

$$E\ (MED_i | x_i) = \phi \cdot \exp\ (x_i \beta_3) - \$5\ , \qquad (3.3b)$$

where

$$\phi = E\ [\exp\ (\varepsilon_{3i})] = \exp\ (\sigma_\varepsilon^2/2) \qquad (3.3c)$$

if the error is normally distributed. Substituting appropriate estimates of β_3 and σ_ε^2 provides an estimate of the expected expense.

While the log expenses are closer to being normally distributed than the untransformed expenses, estimates based on this model yield inconsistent estimates of Eqs. (3.3b) and (3.3c). The existence of

---

[2] If the matrix of explanatory variables is well conditioned and the explanatory power of the equation is low, we would then expect the error term ε_{3i} in Eq. (3.3a) to be more normal than that in Eqs. (3.1) and (3.2). Figure 3.4 is a normal plot of the OLS residuals after fitting Eq. (3.3a).

[3] If the variance is known, then the relative efficiency of the log model to the raw mean is $[\exp\ (\sigma^2) - 1]/\sigma^2$, which is appreciably greater than 1 for the values encountered in the HIS data. If the variance is unknown, the efficiency gain is

$$[\exp\ (\sigma^2) - 1]/(\sigma^2 + \sigma^4/2k)\ ,$$

k being the ratio of the degrees of freedom for estimating σ^2 to the sample size for the group being predicted. Again for our data, there is an efficiency gain from the log model over untransformed analysis.

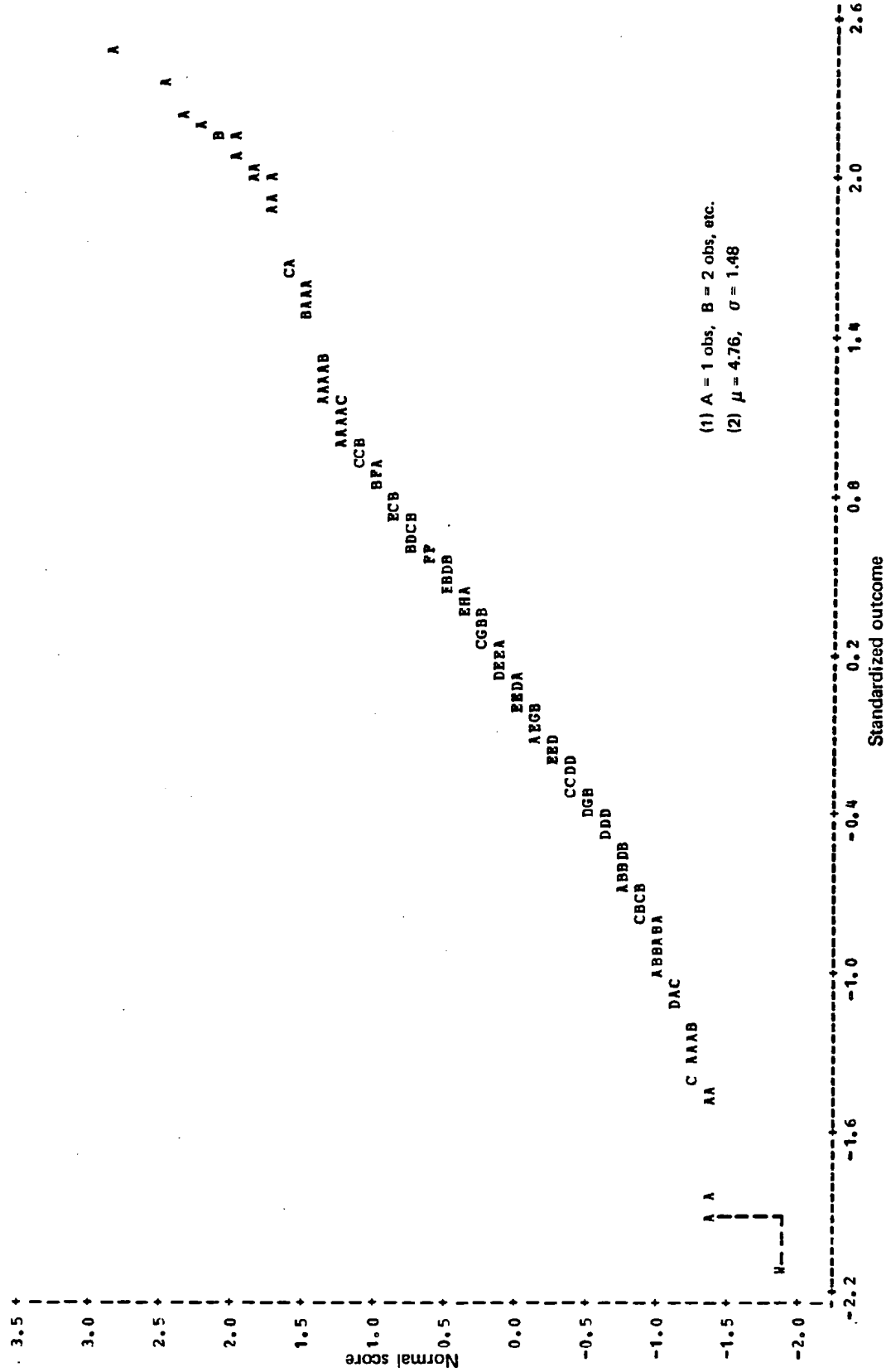


Fig. 3.3--Normal plot of log (MED + $5) for the free plan, Dayton Year 1

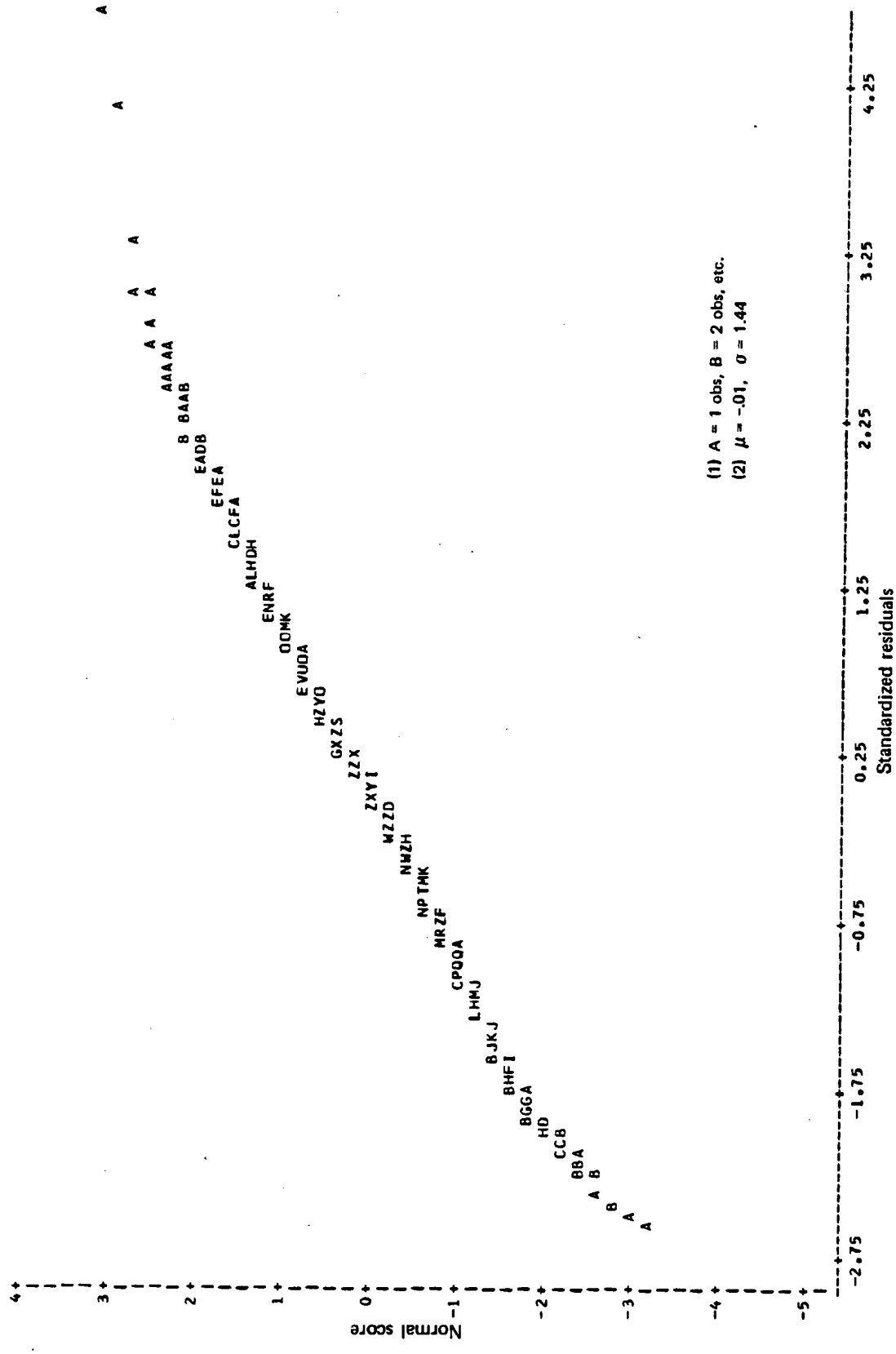


Fig. 3.4--Normal plot of residuals from the one-part model, Dayton Year 1

a large number of nonspenders makes it impossible for the data to be log normal. Moreover, the decision to spend or not (and hence the departure from normality) is systematically related to the covariates. As a result, Eq. (3.3c) does not hold. Instead, the expected value of the exponentiated error is a function of x. Thus, the forecasts based on Eqs. (3.3b) and (3.3c) are statistically inconsistent.

TWO-PART MODEL

The two-part model is an attempt to correct the problem with nonspenders in the one-part model by separating behavior into two stages, first a decision to have positive expenses, and then a decision about the level of expenses, *conditional* on its being positive. More formally, the model has two equations. The first is a probit equation for the dichotomous event of having zero or positive expenses:[4]

$$I_i = x_i\delta_1 + \eta_{1i} , \qquad \eta_{1i} \sim N(0,1) , \tag{3.4a}$$

where MED > 0 if $I \geq 0$, and MED = 0 otherwise. The second equation is a linear model on the log scale for positive expenses:

$$\log (MED_i | I_i > 0) = x_i\delta_2 + \eta_{2i} , \tag{3.4b}$$

where

$$\eta_{2i} \sim N(0,\sigma^2) .$$

[4]A reasonable alternative is a logistic regression. For the range of probabilities in the HIS data, the probit and logistic are very similar. The probit has the advantage that we can examine correlations, both cross-sectionally and intertemporally. These issues are discussed under "Intrafamily Correlation," below.

The expected medical expenditure for an individual with characteristics x_i is

$$E\ (MED_i|x_i) = P_i \cdot E\ (MED_i|MED_i > 0, x_i)$$

$$= P_i \exp\ (x_i\delta_2 + \sigma^2/2)\ , \qquad (3.4c)$$

where

$$P_i = Pr\ (MED_i > 0) = Pr\ (I_i \geq 0) = \Phi(x_i\delta_1)\ .$$

The expected expense in Eq. (3.4c) can be estimated by substituting appropriate estimates of δ_1, δ_2, and σ^2. The estimate of the expected expense provided by Eq. (3.4c) can be statistically inconsistent if the error term η_{2i} in Eq. (3.4b) is not normally distributed.

This model should provide more accurate estimates than the one-part model because it fits the distribution more closely. Compare Figs. 3.4 and 3.5, which are normal plots for the residuals for the two models for the first year of Dayton.

The likelihood function for this model is

$$L(\delta_1,\delta_2,\sigma^2) = \prod_{i=1}^{n} L_i(\delta_1,\delta_2,\sigma^2)\ ,$$

where L_i is the likelihood function for the ith observation. For simplicity, assume that the observations have been sorted in such a way that the first N observations have positive expenses, and the last n - N observations have zero expenses.

For each of the first N observations, we can compute the likelihood function as

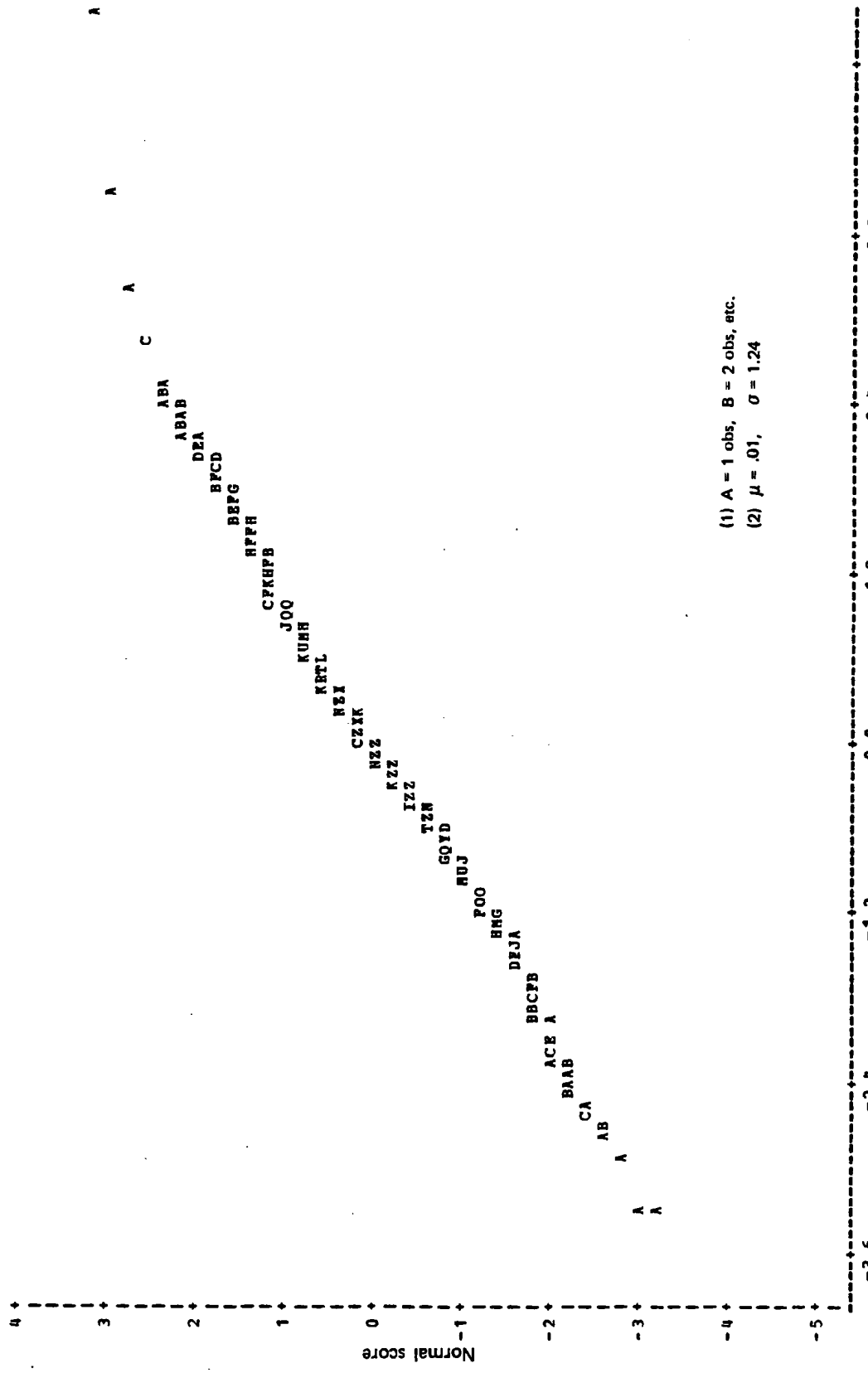


Fig. 3.5--Normal plot of residuals from the expense equation of the two-part model, Dayton Year 1

$$L_i = \Pr\,(MED_i > 0 | x_i) \cdot \text{density}\,(MED_i | MED_i > 0, x_i)$$

$$= \Phi(x_i\delta_1) \cdot \frac{1}{\sigma_{\eta_2}} \phi\left(\frac{y_i - x_i\delta_2}{\sigma_{\eta_2}}\right), \qquad i = 1, \ldots, N ,$$

where ϕ = standard normal p.d.f.,

$y_i = \log\,(MED_i)$.

For each of the last n - N observations, the likelihood function is

$$L_i = \Pr\,(MED_i = 0 | x_i) = 1 - \Phi(x_i\delta_1) , \qquad i = N + 1, \ldots, n .$$

Therefore the likelihood function is

$$L(\delta_1, \delta_2, \sigma^2_{\eta_2}) = \left\{ \prod_{i=1}^{N} \Phi(x_i\delta_1) \cdot \prod_{i=N+1}^{n} [1 - \Phi(x_i\delta_1)] \right\}$$

$$\cdot \left\{ \prod_{i=1}^{N} \frac{1}{\sigma_{\eta_2}} \phi\left(\frac{y_i - x_i\delta_2}{\sigma_{\eta_2}}\right) \right\} . \qquad (3.4d)$$

The important point about L in Eq. (3.4d) is that it factors into two multiplicative terms. The first term,

$$L_1(\delta_1) = \prod_{i=1}^{N} \Phi(x_i\delta_1) \cdot \prod_{i=N+1}^{n} [1 - \Phi(x_i\delta_1)] , \qquad (3.4e)$$

depends *exclusively* on parameters in Eq. (3.4a); the second term,

$$L_2(\delta_2, \sigma^2_{\eta_2}) = \prod_{i=1}^{N} \frac{1}{\sigma_{\eta_2}} \phi\left(\frac{y_i - x_i\delta_2}{\sigma_{\eta_2}}\right) , \qquad (3.4f)$$

depends *exclusively* on parameters in Eq. (3.4b).[5]

Because of the separability, maximizing the likelihood function (3.4d) is equivalent to maximizing the likelihood functions (3.4e) and (3.4f) separately. Therefore, the maximum likelihood estimate for the one-part model can be obtained by joining the maximum likelihood estimate for δ_1 in Eq. (3.4a) and the maximum likelihood estimate for δ_2 and $\sigma^2_{\eta_2}$ in Eq. (3.4b).

Three alternative models--the Tobit, self-selection, and the adjusted Tobit models--have been proposed in the econometric literature for related problems. Appendix A compares the two-part model with these alternatives and explains our preference for the two-part model specification.

FOUR-PART MODEL

The four-part model is an attempt to model the distribution of medical expenses more closely than the one- and two-part models. As our sample size increased after pooling several site-years of data, a departure from log normality in the far-right tail (the very largest expenditures) became evident. The logarithm of positive expenses is too long-tailed (individuals in the right tail spend too much) to be log normal, as indicated in Fig. 3.6. Because such nonnormality occurs more in total medical expenses and far less in ambulatory expenses (see Fig. 3.7), the problem is due to inpatient utilization having both different probabilities and different conditional (log) means and variances.

The four-part model separates the population into three groups: nonspenders, ambulatory-only spenders, and spenders with inpatient (INP) utilization. The four equations are

---

[5]The separability of the likelihood functions is a consequence of the way conditional densities are calculated. It does *not* depend on any independence assumption between Eqs. (3.4a) and (3.4b). Actually, the error terms η_{1i} and η_{2i} in Eqs. (3.4a) and (3.4b) might very well be correlated; the correlation does not affect the separability of the likelihood function in any way.

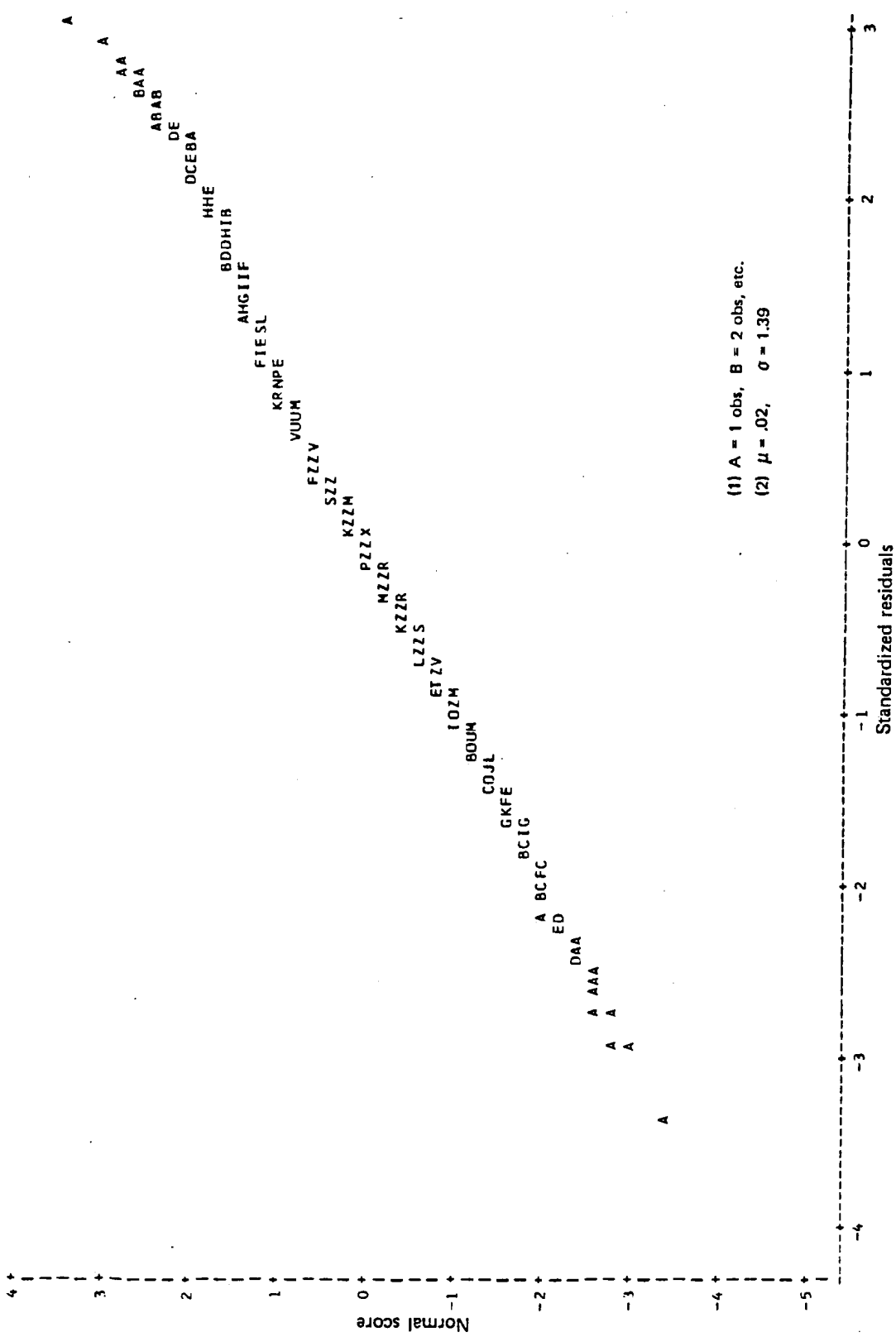


Fig. 3.6--Normal plot of residuals from the expense equation of the two-part model, nine site-years, family deductible plan

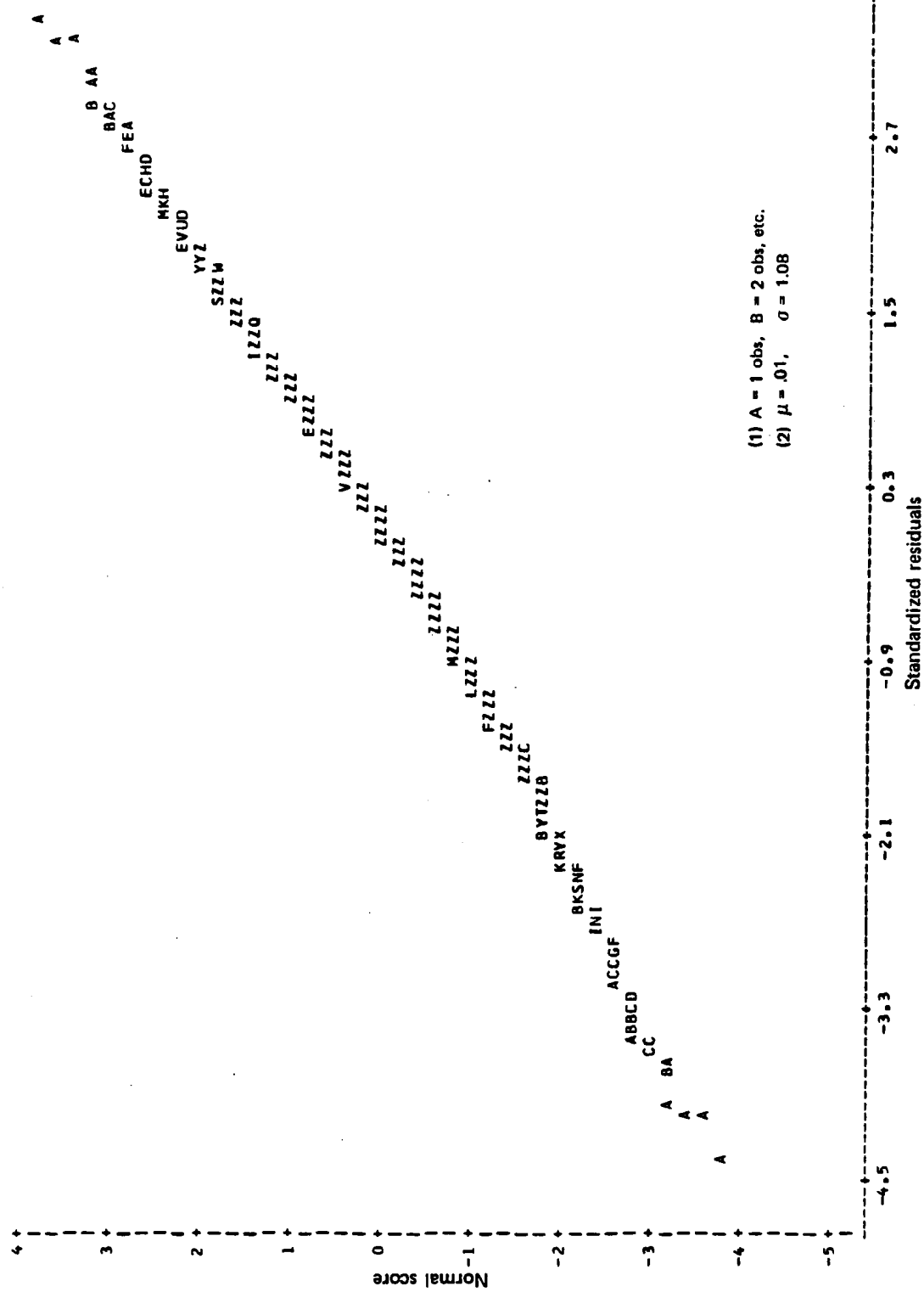


Fig. 3.7--Normal plot of residuals from the ambulatory expense equation, nine site-years

$$\Pr\ (MED_i > 0) = \Phi(x_i\gamma_1)\ , \tag{3.5a}$$

$$\Pr\ (INP_i > 0 \,|\, MED_i > 0) = \Phi(x_i\gamma_2)\ , \tag{3.5b}$$

$$\log\ (MED_i \,|\, MED_i > 0,\ INP_i = 0) = x_i\gamma_3 + \nu_i\ , \tag{3.5c}$$

$$\log\ (MED_i \,|\, INP_i > 0) = x_i\gamma_4 + \omega_i\ . \tag{3.5d}$$

As in the two-part model, the likelihood function for this model is multiplicatively separable in the parameters because of the way conditional densities are calculated.[6] Therefore the maximum likelihood analysis of Eqs. (3.5a) through (3.5d) is to estimate the four equations separately.

The distribution of log medical expenses in Eq. (3.5d) (the subsample with positive inpatient expenses) is appreciably long-tailed, as Fig. 3.8 indicates. To reduce the influence of extreme values, we estimated this equation with Tukey's biweight method, computed by the following iterated weighted least squares method. (Mosteller and Tukey, 1977, Chapter 14.) The OLS estimate is used as starting value. For each iteration, the new weights are derived by

$$w_i = [1 - (r_i/c)^2]^2\ , \quad |r_i| \leq c\ ,$$

$$= 0\ , \quad |r_i| > c\ ,$$

where r_i is the standardized residual, standardized by a robust estimate of scale (.6745 · median abolute deviation). The parameter

---

[6] A plausible alternative model is to use two two-part models; one for ambulatory care and one for inpatient care. Unfortunately, the likelihood function for this model does not have the separability property because of the correlation between ambulatory and inpatient use. In particular, standard errors of the estimates of expected expenses are difficult to calculate.

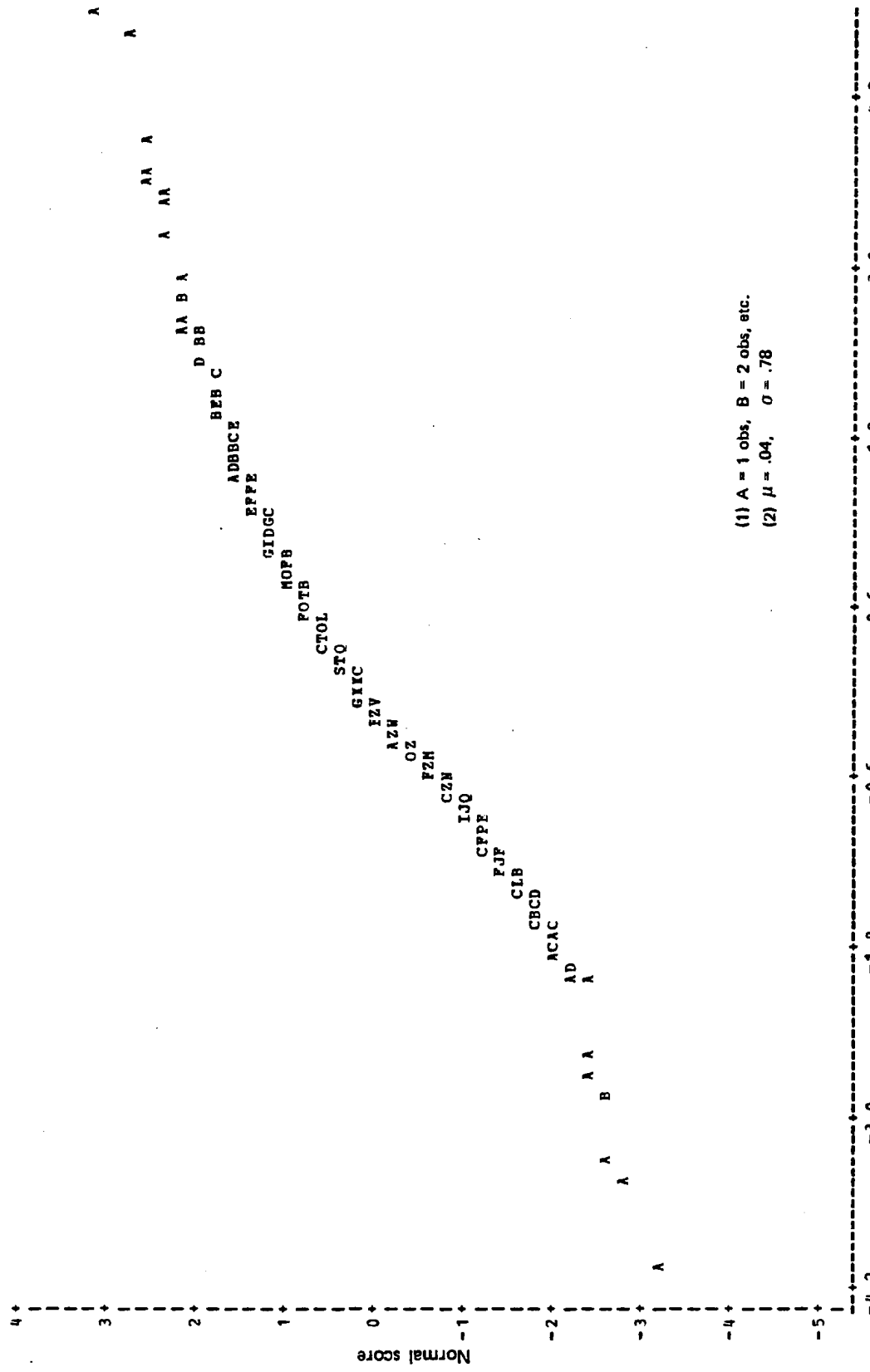


Fig. 3.8--Normal plot of residuals from the fourth equation (positive expense with positive inpatient expense) of the four-part model, nine site-years

c is taken to be 5.[7]

If the error terms ν and ω in Eqs. (3.5c) and (3.5d) were both normal, then the expected medical expense would be

$$P_i[(1 - \pi_i) \exp (x_i\gamma_3 + \sigma_\nu^2/2) + \pi_i \exp (x_i\gamma_4 + \sigma_\omega^2/2)] ,$$

where $P_i = \Pr (MED_i > 0)$,
$\pi_i = \Pr (INP_i > 0 | MED_i > 0)$.

However, the normal assumptions are not satisfied. The distribution of log (MED) for the sample with inpatient expenses [Eq. (3.5d)] is appreciably long-tailed. The distribution of log (MED) for the subsample with ambulatory expenses only [Eq. (3.5c)] is slightly short-tailed, as indicated by Fig. 3.9. As a result of the nonnormality, the normal correction [exp $(\sigma^2/2)$] for the retransformation from the logarithmic scale to the untransformed dollar scale would lead to estimates for the mean expenditure that are statistically inconsistent. In the next section we propose a statistically consistent method to retransform the four-part model's estimates.

SMEARING ESTIMATE

In our discussion of the one-part, two-part, and four-part models above, we have raised the retransformation issue, where inappropriate use of the normal assumption can yield statistically inconsistent predictions of expected expenses. While the development of the transformed models is motivated by approximating the normal assumption as closely as possible, the error distributions, even in the four-part model, still exhibit deviations from the normal assumption.

---

[7]We also explored the use of other values of c, which produced very similar results.

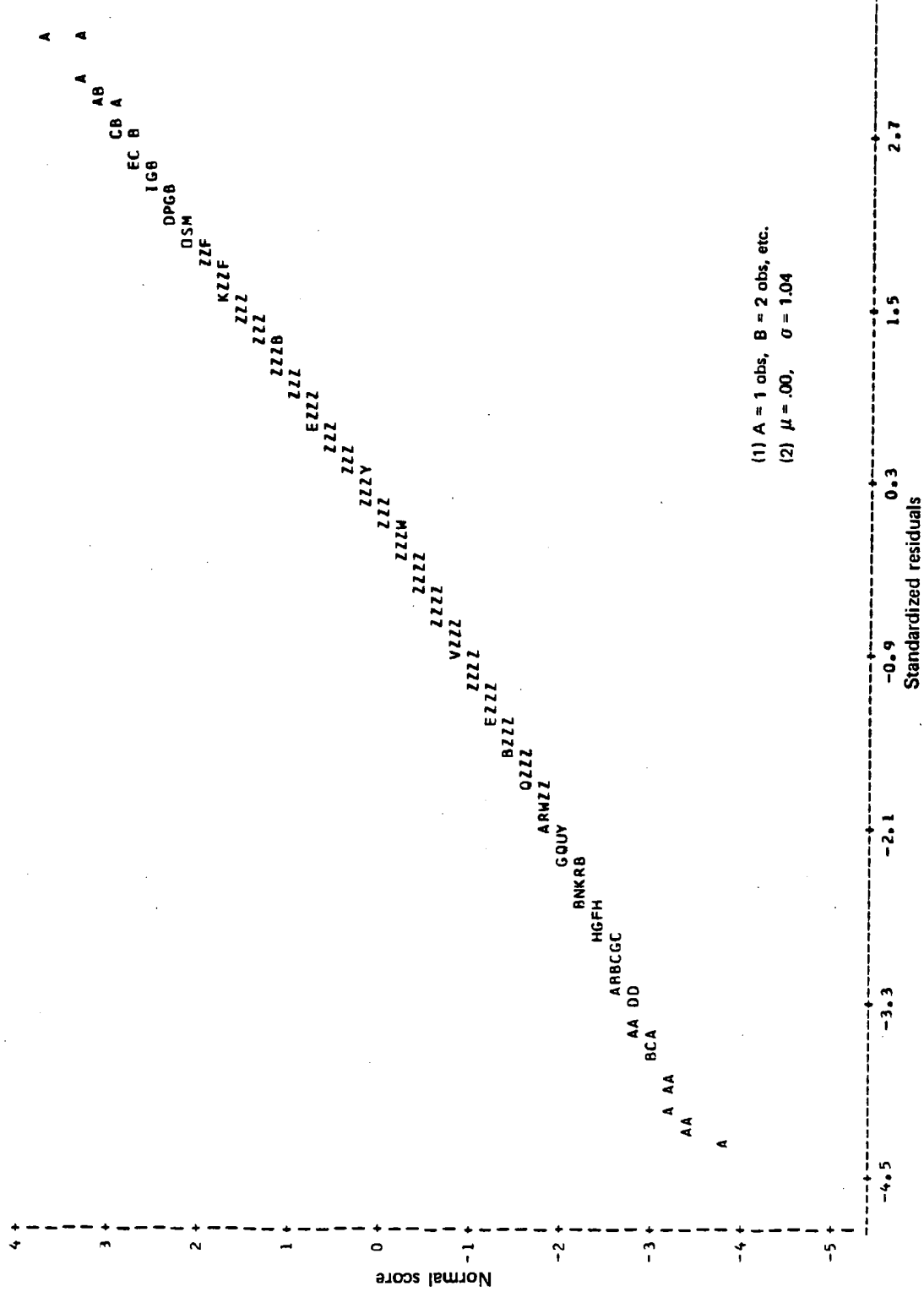


Fig. 3.9--Normal plot of residuals from the third equation (positive expenses with zero inpatient expense) of the four-part model, nine site-years

Without making the normal assumption, the general form of the retransformation bias for a log linear model

$$\log Y = x\beta + \varepsilon \tag{3.6a}$$

is given by

$$E(Y) = e^{x\beta}\phi(x) ,$$

$$\phi(x) = E[\exp(\varepsilon)|x] . \tag{3.6b}$$

If the error distribution does not depend on the characteristics x, the retransformation bias is a constant

$$\phi \equiv E \exp(\varepsilon) . \tag{3.6c}$$

If the error distribution is further assumed to be normal, the retransformation bias is given by

$$\phi = E \exp(\varepsilon) = \exp(\sigma^2/2) , \tag{3.6d}$$

where

$$\sigma^2 = \mathrm{Var}(\varepsilon) .$$

In Appendix B, we develop and discuss a nonparametric estimate, the smearing estimate, of the retransformation factor ϕ. The smearing estimate is the sample average of the exponentiated least squares residuals:

$$\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{\varepsilon}_i) , \tag{3.6e}$$

where n = sample size ,

$\hat{\varepsilon}_i = \log Y_i - x_i\hat{\beta}$,

$\hat{\beta}$ = OLS estimate of β .

The smearing estimate is statistically consistent for the retransformation problem if the error distribution does not depend on the characteristics x. Moreover, when the normal assumption indeed holds, the nonparametric smearing estimate has high efficiency relative to the parametric normal retransformation (3.6d) for a wide range of parameter values.

Unfortunately, the smearing estimate cannot be applied to the one- and two-part models for these data. For both models, the error distribution depends on the characteristics. In the one-part model, the probability of having zero expense depends on most of the observed characteristics. In the two-part model, the probability of having inpatient expense, and therefore large expense, depends on plan, sex, and age. Thus, for both models and our data, the retransformation bias ϕ will not be a constant, but will instead depend on the covariates. The smearing estimate applied to the entire sample is inappropriate in these cases. In principle, we could apply the smearing estimate on subsamples in which the retransformation bias is constant. For example, we could apply the smearing estimate to each plan, sex, and age group separately for the two-part model. However, apart from being cumbersome, the sample size in each cohort is so small that the resulting predictions would be very noisy.

Although neither the smearing estimate nor the normal retransformation [Eq. (3.6d)] is appropriate for the one- and two-part models for these data, the empirical findings presented in Chapter 4 for these models are based on the normal retransformation (3.6d). However, we also estimated the smearing retransformations for these models, and compared the results with those based on the normal retransformation. As Chapter 5 indicates, the normal retransformation results in better predictions than the smearing

method for the one- and two-part models.

For the four-part model, the error distribution's only dependence on the covariates is that the error term ν in Eq. (3.5c), the ambulatory-only expenses, is heteroscedastic by plan. Such heteroscedasticity is to be expected because in the zero coinsurance plan, all participants face zero coinsurance; but in other plans, those participants who exceeded the Maximum Dollar Expenditure face zero coinsurance at the margin, whereas the other participants face the nominal coinsurance rate. As a result, the residual variation should be greatest in the family deductible plan (PFD). Table 3.1 shows the result of an analysis of variance on the residual variation by plan and site-year. As can be seen, the 50-percent (P50) and family deductible (PFD) plans have larger variations than the lower coinsurance plans.

Because of the dependence of the error distribution on plan, the smearing estimate is applied to each plan separately in Eq. (3.5c).

INTRAFAMILY CORRELATION

The errors in our models exhibit a substantial amount of correlation among family members. First, there is a positive correlation among the decisions to receive care [Eqs. (3.4a), (3.5a), and (3.5b)]. Second, there is a positive correlation among the expenses incurred by nonzero spenders in the same family [Eqs. (3.3), (3.4b), (3.5c), and (3.5d)].

Failure to account for these correlations yields inefficient estimates of the coefficients and incorrect estimates of the standard errors. In particular, if the positive correlation is not accounted for, we would underestimate the standard errors for family level (constant within family) variables, such as coinsurance and income effects.

The intrafamily correlations among decisions to have positive expenses are found to be positive and nearly constant across family roles and sizes. The same is true for the intrafamily correlation

Table 3.1

ANOVA OF RESIDUAL VARIATION ON PLAN AND SITE-YEAR[a]

Variable[b]	Coefficient	Standard Error
Intercept	-.955	.062
P25	.080	.057
P50	.257	.056
PFD	.300	.054
IDP	.219	.057
DAY2	.107	.075
DAY3	.132	.067
SEA1	.020	.076
SEA2	.154	.077
FIT1	.139	.094
FIT2	.238	.090
FRA1	.121	.080
FRA2	.094	.083

[a]The dependent variable is log [log (ϕ)] for a given plan and site-year, where ϕ is the smearing estimate in Eq. (3.5c) for each plan and site-year. In other words,

$$\phi = \frac{1}{n} \Sigma \exp(\nu_i) ,$$

where n = sample size for this plan and site-year, ν_i = least squares residuals, and the summation extends over individuals in this plan and site-year. The log log transformation is analogous to using log ($\hat{\sigma}^2/2$) in the normal case, which produces a location-shift specification:

$$\hat{\sigma}^2 = \hat{\sigma}^2\chi^2/\text{d.f.}, \log \hat{\sigma}^2 = \log \sigma^2 + \log \chi^2/\text{d.f.} .$$

Other specifications lead to very similar results.

[b]The free plan in Dayton Year 1 is the omitted group. The site abbreviations DAY, SEA, FIT, and FRA stand for Dayton, Seattle, Fitchburg, and Franklin.

among positive expenses.[8] Thus, the correlations approximate those of a variance-components model with a family specific error component.[9] For nonzero medical expenses in the two-part model, the equation would be

$$\log (MED_{fi} | MED_{fi} > 0) = x_{fi}\beta + \mu_f + \varepsilon_{fi} , \qquad (3.7a)$$

where x_{fi} = a row vector of independent variables for person i in family f,

β = a column vector of coefficients to be estimated,

μ_f = unmeasured family (f) effect,

ε_{fi} = unmeasured individual (i) effect.

Further, we assume that

$$\mu \sim N(0,\sigma_\mu^2) \text{ i.i.d. across families,}^{10} \qquad (3.7b)$$

$$\varepsilon \sim N(0,\sigma_\varepsilon^2) \text{ i.i.d. across persons,} \qquad (3.7c)$$

$$E(\mu X) = E(\varepsilon X) = E(\mu\varepsilon) = 0 . \qquad (3.7d)$$

The covariance matrix of the error vector $\mu + \varepsilon$ is

$$\text{Cov}(\mu + \varepsilon) = \sigma^2(1 - \rho)I_n + \sigma^2\rho D ,$$

---

[8] The one major exception is that the father-child correlation is smaller than the other correlations.

[9] See Balestra and Nerlove (1966), Maddala (1971), Searle (1971, Chapters 9-11), and Mundlak (1978) for a discussion of this model in a regression context.

[10] We cannot estimate the equation with a fixed-effects model because we would be unable to estimate the effects of insurance plan and other family variables. For example, there would be no variation in insurance plan variables because insurance coverage is constant within families.

where $\rho = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_\varepsilon^2)$ is the intrafamily correlation, $\sigma^2 = \sigma_\mu^2 + \sigma_\varepsilon^2$, and D is a block diagonal matrix with a block of 1's for each family, with blocks $i_{m_j} i'_{m_j}$, where i is a column vector of 1's and m_j is the size of the jth family.

The regression coefficients and standard errors for expenditure Eqs. (3.3), (3.4b), (3.5c), and (3.5d) are estimated from this model by maximum likelihood.[11] For each presumed value of intrafamily correlation, the regression coefficients are estimated by generalized least squares. The optimal (maximum likelihood) value of intrafamily correlation is then found by using the Newton-Raphson algorithm.

A similar estimation model was considered for probit Eqs. (3.4a), (3.5a), and (3.5b). However, with the unbalanced design (e.g., unequal family sizes), the computation for the multivariate probit model is prohibitively expensive. Instead, we estimated those equations with univariate probit equations, treating individuals as if they were stochastically independent; we then estimated an upper bound for the standard error. Because the precision associated with the decision to seek care contributes only a small fraction to the overall variance of the prediction for total medical expenditure, we lose little by bounding this standard error. Appendix C contains further details.

In this discussion, we have only addressed the intrafamily correlation within the same equation. There is also the possibility of cross-equation correlation, e.g., between one person's decision to seek care and another family member's decision about the level of care. The models do not include the possibility of such a correlation at this point because these "cross"-correlations are not significantly different from zero in our data.

In the four-part model, we have pooled our observations from different sites and years to estimate inpatient Eqs. (3.5c) and (3.5d). The observations from the same person in different years

---

[11] We are indebted to Dan Relles at Rand, and his statistical package STATLIB, for providing an inexpensive method for handling our unbalanced (e.g., unequal family sizes) design.

are correlated intertemporally. We have estimated an upper bound for the standard errors similar to the adjustment for the intra-family correlation, with the result that the bias in the standard errors due to the intertemporal correlation is at most 10 percent.

Chapter 4
EMPIRICAL FINDINGS

The size and precision of the estimates of the demand response for medical services can be quite sensitive to the estimation model used. Regression with untransformed expenses yields very imprecise results. The use of the logarithmic transformation reduces the imprecision caused by a few large expenditures, but the results vary greatly among the transformed models. We will use the first 2 years in the Dayton, Ohio site to explain the results of our analysis; other sites and years exhibit the same pattern. Because our primary concern is the effect of insurance on demand, we will focus on the differences among the experimental insurance plans.

INFERENCES

ANOVA with untransformed expenses as the dependent variable yields highly imprecise results, even with sample sizes around 1000 for a single site-year. The "ANOVA" column in Tables 4.5 through 4.8 give the results for the first 2 years of Dayton. The plan differences are mostly insignificant and show no meaningful pattern.

ANOCOVA with untransformed expenses also yields highly imprecise estimates, as can be seen in Table 4.1. Very few of the covariates show a statistically significant effect. Many of the coefficients are contrary to intuition (e.g., we would anticipate lower utilization for plans with higher coinsurance rates). Many (12 out of 26) of the coefficients have different signs in the 2 years. Obviously, if we were restricted to the use of ANOCOVA on untransformed expenditures, we would need a much larger sample size to detect any meaningful differences among the insurance plans.

Compared with the untransformed ANOVA and ANOCOVA models, all the transformed models show more significant plan effects that follow

Table 4.1

REGRESSION RESULT FOR UNTRANSFORMED EXPENSES[a]

	Dayton Year 1		Dayton Year 2	
Variable[b]	Coefficient	Standard Error	Coefficient	Standard Error
β_o	1298	895	-445	621
P25	28	125	-41	86
P50	117	140	-195	95
PFD	-137	125	-79	85
IDP	24	174	-185	119
EXAM	109	89	-12	62
WEEKLY	100	89	-46	61
YR3	-137	89	46	61
LINC	-115	89	48	59
LFAM	-15	101	108	66
BLACK	22	160	78	109
AFDC	-562	284	-109	190
NOMD	-40	224	-230	164
NOMDVIS	115	151	63	104
INMDVIS	-298	162	-269	111
HLTHG	193	100	63	69
HLTHFP	508	206	44	143
PAINL	116	101	-23	70
PAINSG	212	206	76	144
WORRL	-97	132	231	91
WORRSG	-48	135	133	94
CHILD	-33	142	-357	97
FEMALE	-275	620	-487	455
FCHILD	183	219	300	164
NEWMEM	242	505	-139	270
MAGE	8	272	337	208
FAGE	192	300	523	247

NOTE: Sample size for Dayton Year 1 is 1110 and for Dayton Year 2 is 1103.

[a]Estimated by ordinary least squares.

[b]See Table 2.2 for definitions.

the expected inverse relationships[1] to coinsurance; they also show greater consistency between Year 1 and Year 2. Moreover, all the transformed models yield similar inferences about other covariate effects. The use of the logarithmic transformation reduces the sensitivity of the results to a few large expenditures both in the ANOVA and the ANOCOVA model for untransformed expenditures.

The estimated regression coefficients for Dayton Years 1 and 2, using the one-part model, are shown in Table 4.2. Most of the important covariates for this model have a significant and meaningful effect on log expenditures. The signs of the estimated coefficients for the covariates are more stable than for ANOCOVA; only six have different signs in the 2 years. The insurance plans exhibit a roughly monotonic pattern. The coefficients for the two-part and four-part models are qualitatively similar to those for the one-part model, as Tables 4.3 and 4.4 indicate.[2] Thus, if we were interested only in qualitative inference, we could accept the results for the one-part model as a satisfactory analysis.

PREDICTIONS

In addition to making inferences about behavior, we have endeavored to predict the cost of alternative health insurance packages. The predictions for medical expenditures based on each of these models are given in Tables 4.5 and 4.6 for Dayton 1 and 2.[3] For each plan, the first column gives the ANOVA prediction, namely, the simple mean, followed by predictions based on ANOCOVA, one-, two-, and four-part models. The prediction standard errors[4] are

---

[1] We expect that as coinsurance increases, expenditure will decrease.

[2] Equation (3.4a) in the four-part model is identical with the probit equation in the two-part model; therefore, it is not repeated.

[3] Results for other sites and years are given in Appendix D.

[4] The standard errors for the ANOVA model are plan-specific. In other words, we have a heteroscedastic ANOVA model. These standard errors are not adjusted for intrafamily correlations. The standard errors for the transformed models are computed by the delta method:

Table 4.2

REGRESSION RESULT FOR THE ONE-PART MODEL[a]

Variable[b]	Dayton Year 1		Dayton Year 2	
	Coefficient	Standard Error	Coefficient	Standard Error
β_o	0.266	1.036	-.262	1.090
P25	-.411	.167	-.455	.170
P50	-.734	.183	-1.002	.186
PFD	-.826	.166	-.711	.169
IDP	-.487	.225	-.821	.230
EXAM	.077	.116	-.065	.119
WEEKLY	.156	.117	-.054	.120
YR3	-.075	.117	.037	.120
LINC	.417	.108	.464	.109
LFAM	-.219	.120	-.187	.120
BLACK	-.375	.208	-.053	.211
AFDC	-.881	.356	-.804	.365
NOMD	-.439	.297	-.843	.322
NOMDVIS	.008	.143	-.183	.161
INMDVIS	-.893	.156	-.812	.175
HLTHG	.175	.110	.266	.119
HLTHFP	.473	.210	.060	.232
PAINL	.110	.107	-.133	.117
PAINSG	.291	.203	.023	.231
WORRL	-.030	.136	.224	.150
WORRSG	.230	.134	.365	.150
CHILD	-.133	.127	-.283	.143
FEMALE	-.075	.569	-.700	.693
FCHILD	-.278	.200	.085	.248
NEWMEM	.981	.467	.277	.413
MAGE	.691	.254	.913	.318
FAGE	.942	.277	1.421	.379

[a]Estimated with a random-effects variance-components model.

[b]See Table 2.2 for definitions.

Table 4.3a

REGRESSION RESULT FOR THE TWO-PART MODEL: PROBIT EQUATION

	Dayton Year 1		Dayton Year 2	
Variable[a]	Coefficient	Standard Error[b]	Coefficient	Standard Error[b]
β_o	-2.387	.987	-3.647	.998
P25	-.646	.164	-.410	.153
P50	-.749	.177	-.740	.162
PFD	-.816	.159	-.648	.146
IDP	-.582	.213	-.789	.195
EXAM	.099	.105	-.058	.100
WEEKLY	.142	.105	.036	.100
YR3	-.112	.104	-.021	.100
LINC	.436	.099	.474	.094
LFAM	-.212	.121	-.327	.111
BLACK	-.534	.169	-.309	.169
AFDC	-.711	.277	-.608	.278
NOMD	-.582	.211	-.666	.229
NOMDVIS	-.155	.151	-.332	.147
INMDVIS	-.809	.186	-.790	.176
HLTHG	-.088	.115	-.027	.112
HLTHFP	.227	.259	-.358	.233
PAINL	.097	.118	-.160	.113
PAINSG	-.091	.252	.045	.253
WORRL	.053	.157	.039	.149
WORRSG	-.059	.163	.111	.162
CHILD	.196	.157	.192	.151
FEMALE	.943	.713	.467	.751
FCHILD	-.593	.252	-.159	.270
NEWMEM	.336	.570	-.143	.453
MAGE	.351	.340	1.137	.388
FAGE	.117	.337	.856	.411

[a]See Table 2.2 for definitions.

[b]These are the estimates of the standard error based on the univariate probit model. They need to be adjusted for intrafamily correlation. For experimental variables (P25, P50, PFD, IDP, EXAM, WEEKLY, and YR3) and the intercept, the appropriate adjustment is to multiply the estimated standard errors by 1.36. See Appendix C for more detail.

Table 4.3b

REGRESSION RESULT FOR THE TWO-PART MODEL: EXPENDITURE EQUATION[a]

	Dayton Year 1		Dayton Year 2	
Variable[b]	Coefficient	Standard Error	Coefficient	Standard Error
β_o	2.305	.944	3.086	1.037
P25	-.197	.142	-.289	.143
P50	-.430	.156	-.648	.160
PFD	-.582	.142	-.380	.143
IDP	-.341	.195	-.472	.205
EXAM	.047	.101	.009	.105
WEEKLY	.086	.102	-.135	.104
YR3	-.033	.101	.096	.103
LINC	.216	.097	.150	.102
LFAM	-.148	.107	-.044	.109
BLACK	-.099	.191	.160	.191
AFDC	.135	.422	-.044	.440
NOMD	.288	.307	-.355	.349
NOMDVIS	.062	.154	.109	.179
INMDVIS	-.555	.156	-.460	.175
HLTHG	.191	.104	.285	.112
HLTHFP	.391	.202	.299	.226
PAINL	.052	.102	-.021	.112
PAINSG	.533	.198	.095	.220
WORRL	-.021	.130	.271	.145
WORRSG	.287	.131	.335	.144
CHILD	-.353	.130	-.582	.148
FEMALE	-.505	.581	-1.312	.695
FCHILD	-.033	.202	.359	.249
NEWMEM	.612	.487	.076	.417
MAGE	.600	.249	.535	.308
FAGE	.950	.286	1.308	.377

[a]Estimated with a random-effects variance-components model.

[b]See Table 2.2 for definitions.

Table 4.4a

RESULTS FOR FOUR-PART MODEL EQ. (3.5b): PROBABILITY OF POSITIVE INPATIENT EXPENSES GIVEN POSITIVE MEDICAL EXPENSES

Variable[a]	Coefficient	Standard Error[b]
β_o	-1.175	.049
AP25	-.116	.067
AP50	-.235	.096
APFD	-.117	.069
AIDP	-.079	.068
FAD[c]	.218	.049
CHILD	-.334	.061

NOTE: The data used to estimate the model were pooled from 9 site-years. The estimates for single site-years were very imprecise.

[a]As noted in Chapter 2, the data exhibit a differential response to plan for adults, but not for children. Other covariates were deleted because the data did not exhibit a linear response in the main-effects-only specification. The data were too thin to estimate the necessary interactions.

[b]The intrafamily correlation for Eq. (3.5b) is negligible. However, the measurements are subject to intertemporal correlation: the same individual is observed repeatedly in different years. Using the methods described in Appendix C, the adjustment for such correlation will increase the standard error by less than 10 percent.

[c]FAD = Female adult.

Table 4.4b

RESULTS FOR FOUR-PART MODEL EQ. (3.5c): LOG MEDICAL EXPENSES FOR POSITIVE-MEDICAL, NO-INPATIENT-EXPENSES SAMPLE[a]

	Dayton Year 1		Dayton Year 2	
Variable	Coefficient	Standard Error	Coefficient	Standard Error
β_o	2.179	.796	2.609	.885
P25	-.282	.121	-.220	.122
P50	-.426	.131	-.564	.135
PFD	-.541	.119	-.406	.123
IDP	-.492	.165	-.333	.171
EXAM	.028	.085	-.005	.089
WEEKLY	.104	.085	-.102	.088
YR3	.032	.085	.068	.087
LINC	.200	.082	.220	.087
LFAM	-.126	.091	-.249	.092
BLACK	-.220	.163	-.018	.168
AFDC	.122	.371	-.171	.385
NOMD	-.026	.265	-.086	.283
NOMDVIS	.004	.130	.063	.149
INMDVIS	-.474	.130	-.394	.148
HLTHG	.090	.086	.212	.095
HLTHFP	.097	.180	.286	.192
PAINL	.073	.086	.095	.095
PAINSG	.506	.177	.287	.183
WORRL	.064	.108	.050	.123
WORRSG	.190	.113	.149	.123
CHILD	-.435	.110	-.278	.122
FEMALE	.038	.505	-.981	.578
FCHILD	.063	.172	.066	.206
NEWMEM	.202	.433	.291	.345
MAGE	.673	.209	.291	.253
FAGE	.638	.245	.951	.313

[a]Estimated with a random-effects variance-components model.

Table 4.4c

RESULTS FOR FOUR-PART MODEL EQ. (3.5d): LOG MEDICAL EXPENSES FOR POSITIVE-INPATIENT-EXPENSE SAMPLE

Variable[a]	Coefficient	Standard Error[b]
β_o	7.444	.081
DAY2[c]	.223	.097
DAY3	.209	.097
SEA1	-.075	.092
SEA2	.157	.097
FIT1	-.024	.106
FIT2	.295	.106
FRA1	.007	.105
FRA2	.138	.109
FAD[c]	-.014	.058
CHILD	-.545	.069

NOTE: The model was estimated with a biweight robust regression using all 9 site-years of data. The site-years were pooled because the estimates for single site-years were very imprecise.

[a]The plan variables were deleted from this specification because they had statistically insignificant coefficients $F(4,735) = 2.03$, and exhibited an erratic pattern. Other covariates were deleted because the data did not exhibit a linear response in the main-effects-only specification. The data were too thin to estimate the necessary interactions.

[b]The standard errors are not corrected for either intrafamily or intertemporal correlation.

[c]DAY2 = Dayton Year 2, etc.
FAD = Female adult.

Table 4.5

DAYTON YEAR 1: AVERAGE PREDICTION AND STANDARD ERROR

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	312.97	314.50	478.30	371.99	414.00
	(35.22)	(86.04)	(62.14)	(40.07)	(32.47)
P25	340.49	342.37	315.52	280.10	311.21
	(67.02)	(90.90)	(42.78)	(32.81)	(28.58)
P50	445.12	431.55	226.99	217.20	265.33
	(236.05)	(109.30)	(35.25)	(29.67)	(30.14)
PFD	179.92	177.65	206.52	183.84	285.73
	(27.24)	(89.25)	(28.01)	(22.01)	(27.22)
IDP	316.22	338.12	291.94	245.46	314.40
	(69.51)	(149.25)	(59.21)	(43.90)	(36.97)

Table 4.6

DAYTON YEAR 2: AVERAGE PREDICTION AND STANDARD ERROR

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	416.96	417.90	584.10	442.78	471.85
	(61.91)	(58.07)	(78.05)	(48.43)	(38.74)
P25	383.87	376.64	368.69	309.88	371.31
	(74.82)	(62.93)	(52.34)	(37.56)	(35.08)
P50	218.67	222.44	211.35	199.69	285.52
	(42.36)	(74.81)	(34.14)	(28.99)	(35.21)
PFD	338.18	339.12	284.30	267.57	345.21
	(66.85)	(61.28)	(40.11)	(33.30)	(33.77)
IDP	227.02	233.32	254.12	234.74	332.56
	(62.89)	(103.02)	(53.72)	(46.52)	(41.41)

given in parentheses. Following the predictions, the plan relatives, i.e., the mean predicted expenditure expressed as percentages of free plan mean predicted expenditures, are given as Tables 4.7 and 4.8 for Dayton 1 and 2. The absolute t-values given in parentheses are based on differences from the free plan as measured on the dollar scale, *not* in proportional terms.[5]

Plan predictions in the various models can depend on the distribution of noninsurance variables, as well as on plan response, because the plans are not perfectly balanced. For a meaningful comparison among the plans, we should correct for any possible differences among the noninsurance characteristics of each plan's enrollees.[6] Therefore, we estimated predicted values for all the participants in the specific site-year, assuming that all of them were assigned to the plan being predicted, and then averaged them.[7] (ANOVA can be done only on actual enrollees.)

$$\text{Var (prediction)} = \text{Var } f(\hat{\beta}_1, \hat{\beta}_2, \ldots)$$

$$\approx \left(\frac{\partial f}{\partial \beta_1}\right)' \cdot \text{Cov } (\hat{\beta}_1)\left(\frac{\partial f}{\partial \beta_1}\right)$$

$$+ \left(\frac{\partial f}{\partial \beta_2}\right)' \cdot \text{Cov } (\hat{\beta}_2)\left(\frac{\partial f}{\partial \beta_2}\right) + \ldots .$$

In the multipart models, the separability of the likelihood function discussed in Chapter 3 implies that the estimated coefficients in different equations are asymptotically uncorrelated.

[5]It should be noted that the predictions on different plans are *correlated*. The standard error for plan differences cannot be computed as if the predictions were independent, because the different predictions are based on other shared covariates as well as plan. The correlation has been accounted for in Tables 4.7 and 4.8, and in Appendix E, which has the results for other site-years.

[6]Although the experiment was designed to be reasonably balanced, perfect balance is unachievable.

[7]The average predictions for actual enrollees are very similar to the predictions for all participants, confirming that the plan assignments were nearly balanced.

Table 4.7

PLAN RELATIVES FOR DAYTON YEAR 1: EXPENDITURES EXPRESSED AS A PERCENTAGE OF THE FREE PLAN[a]

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	100 (-)	100 (-)	100 (-)	100 (-)	100 (-)
P25	109 (0.36)	109 (0.22)	66 (2.37)	75 (1.93)	75 (3.36)
P50	142 (0.55)	137 (0.84)	47 (3.77)	58 (3.34)	64 (4.32)
PFD	57 (2.99)	56 (1.09)	43 (4.30)	49 (4.46)	69 (4.30)
IDP	101 (0.04)	108 (0.14)	61 (2.30)	66 (2.24)	76 (2.63)

[a]Absolute t-values given in parentheses are based on the difference between that plan and the free plan as measured on the dollar scale, not in proportions.

MODEL COMPARISONS

The predictions by plan vary greatly according to the model used. The simple ANOVA and ANOCOVA on untransformed expenditures yield very noisy results that do not have the monotonic response to insurance that one would expect. The one-part model produces monotonic and more precise results. It also exhibits a more pronounced response to insurance plan than do any of the other approaches. The two-part model yields smaller estimates of plan differences than the one-part model. However, the plan relatives of the two-part model are lower for the pay plans than one would expect from the unbiased ANOVA and ANOCOVA on untransformed expenditures. Among a total of 34 site-year-plan specific predicted plan relatives in Tables 4.7 and 4.8 and in Appendix E, the one-part model "underestimates" 28 of them and the two-part model "underestimates" 21, both compared with the unbiased ANOVA results.

Table 4.8

PLAN RELATIVES FOR DAYTON YEAR 2: EXPENDITURES EXPRESSED AS A PERCENTAGE OF THE FREE PLAN[a]

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	100 (-)	100 (-)	100 (-)	100 (-)	100 (-)
P25	92 (0.34)	90 (0.48)	63 (2.57)	70 (2.40)	79 (2.77)
P50	52 (2.64)	53 (2.04)	36 (4.71)	45 (4.65)	61 (4.59)
PFD	81 (0.86)	81 (0.93)	49 (3.78)	60 (3.30)	73 (3.50)
IDP	54 (2.15)	56 (1.55)	44 (3.74)	53 (3.28)	70 (3.20)

[a]Absolute t-values given in parentheses are based on the difference between that plan and the free plan as measured on a dollar scale, not in proportions.

The four-part model adjusts explicitly for the nonnormality in the right tail and the heteroscedasticity in log positive ambulatory expenses. This model also corrects for the fact that the inpatient utilization (both probability and level) is less responsive to insurance plan than is the ambulatory utilization. (Without separating the differential responsiveness in the inpatient and ambulatory utilization, the one- and two-part models actually concentrate on the median expenses, which are largely determined by ambulatory utilization. Therefore, the plan relatives predicted from these models are really plan relatives for the more responsive ambulatory utilization.) As a result, its predicted plan relatives are smaller in magnitude than those for the one- and two-part models. Among the 34 plan relatives, the four-part model "underestimates" 14 of the plan relatives compared with the ANOVA results.

Chapter 5
SPLIT-SAMPLE ANALYSIS

With several plausible models as candidates, we must choose the most appropriate model. The task is especially important because the various models produce very different predictions. The development of the transformed models described in Chapter 3 was mainly based on consistency considerations. For very large sample sizes, when the asymptotic bias dominates other components of error, the four-part model will be better than the other transformed models because it is free from the inconsistency in the one- and two-part models. However, for a fixed finite sample size, a model that admits a small amount of bias to achieve higher precision (smaller variance) can outperform a model free from bias at the expense of lower precision. The models should therefore be evaluated based on a compromise between bias and precision.

By using more complicated models to fit the HIS data, we run a substantial risk of overfitting the data: these additional complications in the models may be simply fitting noise in the data. To ensure that the criteria can detect overfitting, we use a split-sample technique, i.e., we estimate the parameters of the model on one-half of the sample and then make forecasts to the other half. The models are evaluated in terms of mean squared forecast error and mean forecast bias on the forecast sample. A model can perform poorly on the forecast sample if it is imprecise, or inconsistent (e.g., due to overfitting the estimation sample).

The more complicated models perform significantly better than the simpler ones in terms of mean squared forecast error and somewhat better in terms of mean forecast bias. Given our present sample size, we cannot reject the hypothesis that the two- and four-part models are equally good. When the models are compared for overfitting, the four-part model behaves better than the other models, and the untransformed models behave worse than the transformed models.

At the end of this chapter, we compare empirically the retransformation methods. The data support the choice of the normal retransformation for the one- and two-part models, and smearing for the four-part model. The choice between homoscedastic and heteroscedastic retransformations is inconclusive.

METHODOLOGY

In choosing between the models, we cannot rely on conventional likelihood ratio tests because our models are not nested in the usual sense. Instead, we have adopted a split-sample approach to evaluate them. The method can be viewed as an application of the classical cross-validation technique, such as McCarthy (1976). Each site-year of data is randomly split into two subsamples--an estimation subsample and a forecast subsample.[1] From the estimation subsample, we derive estimates of the parameters (regression coefficients, variances, and smearing coefficients) for each of the models. We then forecast the expenditures for each person in the forecast subsample, using the models fitted on the estimation subsample. The forecasts are then compared with the actual medical expenditures observed.

One advantage of this approach is that it guards against overfitting the data. Some models are more complex than others. The additional features of the more complicated models might be merely fitting noise in the data used for estimation. If overfitting occurs, the forecasts to a new data set will perform poorly.

The specific criteria that we use to evaluate the forecasts are based on the mean forecast bias (MFB) and the mean squared forecast error (MSFE)

$$MFB = \frac{1}{m} \sum_{i=1}^{m} (\hat{MED}_i - MED_i) , \qquad (5.1)$$

[1] In fact we have done this twice, producing two independent random splits, labeled A and B in Tables 5.1, 5.3, and 5.4. With the second split, we can check our results from the first to reduce the chance that they were due to the luck-of-the-draw.

$$MSFE = \frac{1}{m} \sum_{i=1}^{m} (\hat{MED}_i - MED_i)^2 , \tag{5.2}$$

where the summation extends over the m individuals in the forecast sample, $\hat{MED}_i$ is the forecast for the ith individual, and MED_i is his actual expense. Table 5.1 gives the forecast sample values for each model. As their names imply, the MFB is a measure of the inconsistency in the forecasts, whereas the MSFE is a measure of the inaccuracy in the forecast. The two measures can be reexpressed as

$$MFB = \frac{1}{m} \sum_{i=1}^{m} [\hat{MED}_i - E(MED_i)] - \frac{1}{m} \sum_{i=1}^{m} [MED_i - E(MED_i)], \tag{5.3a}$$

$$MSFE = \frac{1}{m} \sum_{i=1}^{m} [\hat{MED}_i - E(MED_i)]^2 + \frac{1}{m} \sum_{i=1}^{m} [MED_i - E(MED_i)]^2 - \frac{2}{m} \sum_{i=1}^{m} [\hat{MED}_i - E(MED_i)] \cdot [MED_i - E(MED_i)] . \tag{5.3b}$$

As Eqs. (5.3a) and (5.3b) indicate, each measure can be expressed as a sum of deficiency in the fitted model (the first term on the right-hand side), and of measurement error (the second term on the right-hand side). In the case of MSFE, there is an additional term, which is the cross-product of the deficiency in the fitted model and measurement error. Given the estimation sample, the conditional expectations of the two measures are

$$E(MFB) = \frac{1}{m} \sum_{i=1}^{m} [\hat{MED}_i - E(MED_i)] , \tag{5.4a}$$

$$E(MSFE) = \frac{1}{m} \sum_{i=1}^{m} [\hat{MED}_i - E(MED_i)]^2 + \frac{1}{m} \sum_{i=1}^{m} Var(MED_i) . \tag{5.4b}$$

Table 5.1

FORECAST SAMPLE VALUES: MEAN FORECAST BIAS AND MEAN SQUARED FORECAST ERROR FOR TWO SPLIT-SAMPLES[a]

(In dollars)

Model	MFB		$\sqrt{\text{MSFE}}$		Modified $\sqrt{\text{MSFE}}$[b]
	A	B	A	B	
Four-part	-14.40	(-6.67)	1524.68	(1455.36)	1385.79
Two-part	-52.19	(-45.22)	1520.59	(1453.11)	1386.71
One-part	+12.27	(18.40)	1515.37	(1455.63)	1400.69
ANOCOVA	-27.36	(-16.17)	1539.24	(1485.10)	1401.20
ANOVA	-33.15	(-6.42)	1543.10	(1475.45)	1398.21

[a]The two sets of results are based on two independently drawn replicate split-samples from the same population.

[b]The modified MSFE is computed after deleting two individuals in the first (A) forecast sample who had large expenditures and hence have a very large influence on MSFE.

Thus, the MFB is an unbiased estimate of the deficiency (bias) in the fitted model. However, the MSFE is a biased estimate of the deficiency (MSE) in the fitted model because of measurement error variance--the second term on the right-hand side.

The actual measures that we use to compare the models are the *differences* in the MFB and the MSFE. For two competing models--say, Models 1 and 2--the conditional expectations of the differences are

$$E\,[MFB(1) - MFB(2)] = \frac{1}{m}\sum_{i=1}^{m}[\hat{MED}_i(1) - E\,(MED_i)]$$

$$- \frac{1}{m}\sum_{i=1}^{m}[\hat{MED}_i(2) - E\,(MED_i)]\,, \quad (5.5a)$$

$$E\,[MSFE(1) - MSFE(2)] = \frac{1}{m}\sum_{i=1}^{m}[\hat{MED}_i(1) - E\,(MED_i)]^2$$

$$- \frac{1}{m}\sum_{i=1}^{m}[\hat{MED}_i(2) - E\,(MED_i)]^2 \,. \qquad (5.5b)$$

In contrast to the measures for each model separately, *both* the MFB difference and the MSFE difference are unbiased estimates of the corresponding differences in the deficiency in the fitted models. Taking the differences removes the measurement error variance term from the MSFE.

Conditioned on the estimation sample, the MFB difference is a constant, whereas the MSFE difference is a random variable. A normal score test for the difference can be developed for the MSFE difference, conditioned on the estimation sample. However, we rejected the normal score test approach because it is extremely unstable. The inclusion or exclusion of a few extreme values, such as catastrophic expenses, can alter the variance of MSFE(1) - MSFE(2) by as much as an order of magnitude.[2] A comparison of MSFE for sample A and the modified MSFE in Table 5.1 also shows how sensitive the MSFE measures are.

Rather than a normal score test, we have decided to use a subpopulation sign test to detect consistent patterns in MSFE.[3] If one model forecasts expenditures appreciably better than another, we expect the pattern to hold consistently across the subpopulations.

---

[2]The conditional variance of the MSFE difference is

$$Var\,[MSFE(1) - MSFE(2)] = \frac{4}{m^2}\sum_{i=1}^{m}[\hat{MED}_i(1) - \hat{MED}_i(2)]^2 \cdot Var\,(MED_i) \,.$$

[3]The subpopulation sign tests for MFB are also presented in the tables. However, such a test does not have the same distributional properties. Conditioned on the estimation sample, MFB is constant.

Therefore, the subpopulation sign test counts the number of subpopulations for which one model performs better than the other in terms of MSFE. Conditioned on the estimation sample (and therefore conditioned on the fitted models), the count follows a binomial distribution, with probability 0.5 and sample size equal to the number of subpopulations under the null hypothesis of no difference between the two fitted models. (Conditioned on the estimation sample, the counts in distinct subpopulations are stochastically independent.) Significantly high or low counts indicate the existence of a consistent pattern, which we take as evidence that one model is significantly better or worse than the other.[4]

For this analysis, the data can be naturally grouped into 43 subpopulations, one for each site, year, and plan combination. Table 5.2 contains the P-values for the null hypothesis with probability 0.5 and sample size 43.

RESULTS: MODEL COMPARISON

The more complicated models perform significantly better than the simpler ones in terms of MSFE. As Table 5.3 indicates, the models are significantly ordered as follows: the four-part model and the two-part model are better than the one-part model and ANOVA, which are better than ANOCOVA. The difference between the four-part model and the two-part model, as well as the difference between the one-part model and ANOVA, is insignificant.

The results for the MFB follow a similar pattern as for MSFE, but on the whole they are less "significant,"[5] as Table 5.4 indicates.

Based on this analysis, we cannot reject the null hypothesis that the two- and four-part models are equally good. On the one hand, the inconsistency in the two-part model is not large enough to

---

[4]The subpopulation sign test tests the null hypothesis median [MSFE(1) - MSFE(2)] = 0, instead of the hypothesis E [MSFE(1) - MSFE(2)] = 0. However, under the stronger null hypothesis that MSFE(1) and MSFE(2) follow the same distribution, the difference is symmetrically distributed, and the median coincides with the expectation.

[5]The binomial distribution does not apply for MFB.

Table 5.2

BINOMIAL TABLE[a]

(n = 43, p = 0.5)

k	Prob (count > k)
22	.500
23	.380
24	.271
25	.180
26	.111
27	.063
28	.033
29	.016
30	.007
31	.003
32	.001

[a]*Table of Binomial Distribution,* National Bureau of Standards Applied Mathematics Series 6, 1950. (Reprinted with correction 1952.)

make the model significantly worse than the four-part model in terms of our criteria. On the other hand, there is not significant evidence that the more complex four-part model overfits the estimation subsample.

However, the split-sample evaluation in terms of MSFE is inherently more favorable to the two-part model than to the four-part model. The size of the estimation subsample is only half that of the actual sample. Therefore the comparison is based on the compromise between bias and precision for the half-sample size instead of the total sample size. With a larger sample size, such as the total sample size, the bias will be more important than it is in the smaller half-sample size, to the disadvantage of the two-part model relative to the four-part model.

RESULTS: OVERFITTING

One of the main reasons for conducting this split-sample analysis was to determine whether certain models overfitted the data.

Table 5.3

SUBPOPULATION SIGN TESTS FOR MEAN SQUARED FORECAST ERROR ON TWO FORECAST SAMPLES[a]

Model 1	Model 2	Number Where Model 1 Has Lower MSFE[b] Than Model 2	
		A[c]	B[d]
Four-part	Two-part	21	24
	One-part	28*	29*
	ANOCOVA	32*	35*
	ANOVA	31*	31*
Two-part	One-part	31*	27
	ANOCOVA	31*	37*
	ANOVA	25	28*
One-part	ANOCOVA	28*	29*
	ANOVA	23	23
ANOCOVA	ANOVA	16	15*

[a]Out of a possible total of 43 site-year plans. The two counts are based on two independently drawn replicate split-samples from the same population.

[b]The subpopulation sign test is not necessarily transitive.

[c]Based on the first randomly drawn split-sample "A."

[d]Based on the second randomly drawn split-sample "B."

*Significant at the 0.05 level (one-sided).

Table 5.4

SUBPOPULATION SIGN TESTS FOR MEAN FORECAST BIAS ON TWO FORECAST SAMPLES[a]

Model 1	Model 2	Number Where Model 1 Has Lower MFB[b] Than Model 2	
		A[c]	B[d]
Four-part	Two-part	21	26
	One-part	28*	27
	ANOCOVA	25	29*
	ANOVA	25	25
Two-part	One-part	23	17
	ANOCOVA	26	30*
	ANOVA	22	29*
One-part	ANOCOVA	21	28*
	ANOVA	18	23
ANOCOVA	ANOVA	15*	15*

[a]Out of a possible total of 43 site-year plans. The two counts are based on two independently drawn replicate split-samples from the same population.

[b]The subpopulation sign test is not necessarily transitive.

[c]Based on the first randomly drawn split-sample "A."

[d]Based on the second randomly drawn split-sample "B."

*Significant at 0.05 level (one-sided) under the binomial distribution. (Does not actually apply to MFB.)

One way to detect overfitting is to see whether a model fits the estimation subsample appreciably better than the forecast subsample. A model wins in Table 5.5 if it performs better than another model on the estimation sample. Holding as fixed the comparison of the two models on the forecast sample (Table 5.3), if one model performs appreciably better than another on the estimation sample than on the forecast sample, then this model overfits the estimation sample more than the other.

A comparison of the subpopulation sign tests in Tables 5.3 and 5.5 indicates that the ANOVA and four-part models exhibit less overfitting than the others. ANOCOVA exhibits the most overfitting. On the estimation sample, ANOCOVA is appreciably better than ANOVA and the one-part model, and somewhat better than the four-part model. In contrast, on the forecast sample, ANOCOVA is significantly worse than all the other models. The one-part model exhibits possible overfitting when compared with the four-part model. The two-part model also overfits the estimation sample when compared with ANOVA and the four-part model.

Our original concern that the four-part model would overfit the data more than the less-complicated models appears to be unfounded. It appears that it is the *form of the model* rather than the *number of parameters* that is crucial. The model that more accurately describes the distribution of medical expenses (the four-part model) exhibits the least overfitting. The model that ignores the distribution of medical expenses (the ANOCOVA model) exhibits the most overfitting.[6]

RETRANSFORMATION METHODS

As a part of the model comparison, we must also validate the choice of the retransformation method. The discussion in Chapter 3 and the preceding discussion in this chapter assume that each of the

[6]ANOVA is also relatively free of overfitting. However, this accomplishment comes at the cost of all the information on use by age, sex, race, income, and other covariates.

Table 5.5

SUBPOPULATION SIGN TESTS FOR MEAN SQUARED FORECAST ERROR ON ESTIMATION SAMPLE[a]

Model 1	Model 2	Number Where Model 1 Has Lower MSFE[b] Than Model 2	
		A[c]	B[d]
Four-part	Two-part	10*	12*
	One-part	20	23
	ANOCOVA	18	21
	ANOVA	29*	31*
Two-part	One-part	27	30*
	ANOCOVA	22	26
	ANOVA	35*	31*
One-part	ANOCOVA	15*	19
	ANOVA	22	24
ANOCOVA	ANOVA	27	25

[a]Out of a possible total of 43 site-year plans. The two counts are based on two independently drawn replicate split-samples from the same population.

[b]The subpopulation sign test is not necessarily transitive.

[c]Based on the first randomly drawn split-sample "A."

[d]Based on the second randomly drawn split-sample "B."

*Significant at the 0.05 level (one-sided) under the binomial distribution. (Does not actually apply to the estimation sample.)

transformed models will have a specific retransformation method--normal homoscedastic for the one-part and two-part models, and homoscedastic smearing for the inpatient and heteroscedastic[7] smearing for the ambulatory-only expenses for the four-part model. In this section, we provide an empirical evaluation of alternative retransformation methods for the one-part, two-part, and four-part models, using the same estimation and forecast subsamples as those used in the previous sections. We report the results for the first (A) replicate; the second (B) replicate yields qualitatively similar results.

We consider four alternative retransformation methods:

1. Normal with homoscedastic error within a site-year (NHOM).
2. Normal with heteroscedastic error by plan within a site-year (NHET).
3. Smearing with homoscedastic error within a site-year (SHOM).
4. Smearing with heteroscedastic error by plan within a site-year (SHET).

In the four-part model, the retransformation for the positive inpatient subsample is always a smearing estimate that is homoscedastic across site-year plans;[8] the four alternative methods apply just to the ambulatory-only expenses. As an illustration, Table 5.6 gives these coefficients for the ambulatory-only equation (3.5c) in the four-part model.

---

[7] We are using the terms "homoscedastic" and "heteroscedastic" in the strong sense when referring to the smearing estimator. The error distributions are treated as heteroscedastic if they are not the same in any sense, which can be either a difference in scale or in shape. In particular, they are heteroscedastic if the retransformation biases $E\, e^{\varepsilon}$ are different. Even when the error distributions are homoscedastic in the weaker sense (Var $\varepsilon_i \equiv$ constant), they can still be heteroscedastic in the strong sense if they have different shapes.

[8] The inpatient error distribution in the inpatient expenses is clearly not normal (see Fig. 3.8); the positive inpatient sample is too small to allow precise estimation of heteroscedasticity.

Table 5.6

LOG NORMAL AND SMEARING RETRANSFORMATION FOR EQ. (3.5c) IN THE FOUR-PART MODEL[a]

Site-Year	Plan	NHOM[b]	NHET[c]	SHOM[d]	SHET[e]
Dayton 1	Free	1.76	1.53	1.60	1.43
Dayton 1	P25	1.76	1.62	1.60	1.47
Dayton 1	P50	1.76	1.74	1.60	1.52
Dayton 1	FD	1.76	2.15	1.60	1.84
Dayton 1	ID	1.76	2.16	1.60	2.04
Dayton 2	Free	1.82	1.66	1.62	1.49
Dayton 2	P25	1.82	1.65	1.62	1.52
Dayton 2	P50	1.82	1.98	1.62	1.77
Dayton 2	FD	1.82	2.35	1.62	1.87
Dayton 2	ID	1.82	1.68	1.62	1.49
Dayton 3	Free	1.88	1.76	1.73	1.60
Dayton 3	P25	1.88	1.83	1.73	1.71
Dayton 3	P50	1.88	1.81	1.73	1.65
Dayton 3	FD	1.88	2.22	1.73	1.98
Dayton 3	ID	1.88	1.79	1.73	1.70
Seattle 1	Free	1.65	1.52	1.48	1.31
Seattle 1	P25	1.65	1.62	1.48	1.50
Seattle 1	P50	1.65	(f)	1.48	(f)
Seattle 1	FD	1.65	1.68	1.48	1.52
Seattle 1	ID	1.65	1.88	1.48	1.65
Seattle 2	Free	1.85	1.74	1.64	1.53
Seattle 2	P25	1.85	1.65	1.64	1.50
Seattle 2	P50	1.85	(f)	1.64	(f)
Seattle 2	FD	1.85	2.07	1.64	1.71
Seattle 2	ID	1.85	2.16	1.64	1.95
Fitchburg 1	Free	1.87	1.76	1.81	1.63
Fitchburg 1	P25	1.87	1.70	1.81	1.61
Fitchburg 1	P50	1.87	2.02	1.81	1.61
Fitchburg 1	FD	1.87	2.33	1.81	2.12
Fitchburg 1	ID	1.87	1.99	1.81	2.20
Fitchburg 2	Free	1.96	1.96	1.69	1.54
Fitchburg 2	P25	1.96	1.82	1.69	1.60
Fitchburg 2	P50	1.96	1.40	1.69	1.33
Fitchburg 2	FD	1.96	2.36	1.69	2.29
Fitchburg 2	ID	1.96	2.09	1.69	1.80
Franklin 1	Free	1.64	1.49	1.57	1.45
Franklin 1	P25	1.64	1.73	1.57	1.67
Franklin 1	P50	1.64	2.26	1.57	2.21
Franklin 1	FD	1.64	1.72	1.57	1.53
Franklin 1	ID	1.64	1.68	1.57	1.56
Franklin 2	Free	1.80	1.79	1.67	1.64
Franklin 2	P25	1.80	1.81	1.67	1.64
Franklin 2	P50	1.80	1.94	1.67	1.97
Franklin 2	FD	1.80	2.09	1.67	1.85
Franklin 2	ID	1.80	1.66	1.67	1.51

[a] The subsample with positive ambulatory expense and no inpatient expense from the B replicate estimation subsample described earlier.

[b] Homoscedastic lognormal retransformation exp ($\sigma^2/2$).

[c] Heteroscedastic (by plan) lognormal retransformation exp ($\sigma^2_{plan}/2$).

[d] Homoscedastic smearing retransformation ϕ = average of exp (ε_i).

[e] Heteroscedastic smearing retransformation ϕ_{plan} = average of exp (ε_i) within plan.

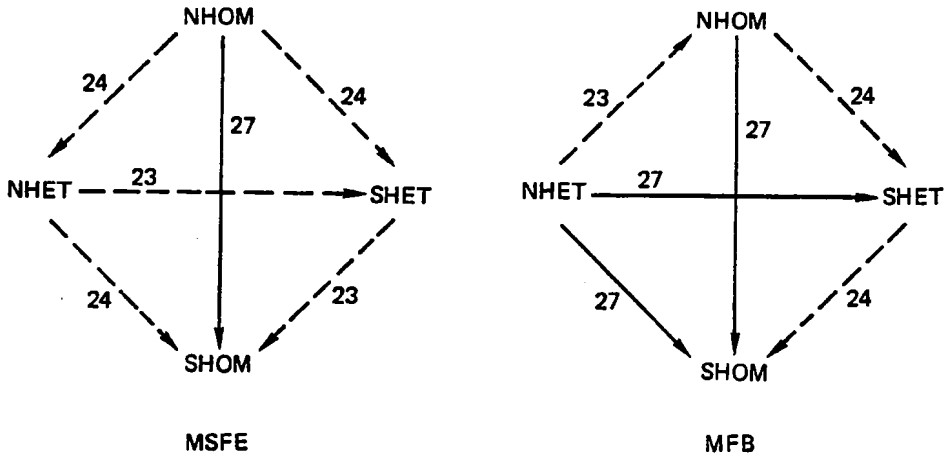
[f] P50 does not exist in Seattle.

In the one-part model, the subpopulation sign test provides support for the choice of normal over the smearing estimates. As Fig. 5.1 indicates, NHOM is significantly better than SHOM in terms of both MSFE and MFB. NHET is "significantly" better than SHET in terms of MFB and weakly better in terms of MSFE. The choice of homoscedastic over heteroscedastic method is inconclusive. Hence, we should choose either NHOM or NHET, but we cannot distinguish between them.

For the two-part model, the split-sample analysis provides similar evidence in the choice of retransformation method. As Fig. 5.2 indicates, the normal retransformation methods dominate the smearing methods. NHET is significantly better than SHET and NHOM is weakly better than SHOM for both criteria, MSFE and MFB. Again, the homoscedastic versus heteroscedastic choice is unclear. However, NHOM weakly dominates SHET and NHET dominates SHOM in terms of both MSFE and MFB. Hence, we should choose either NHOM or NHET, but we cannot distinguish between them.

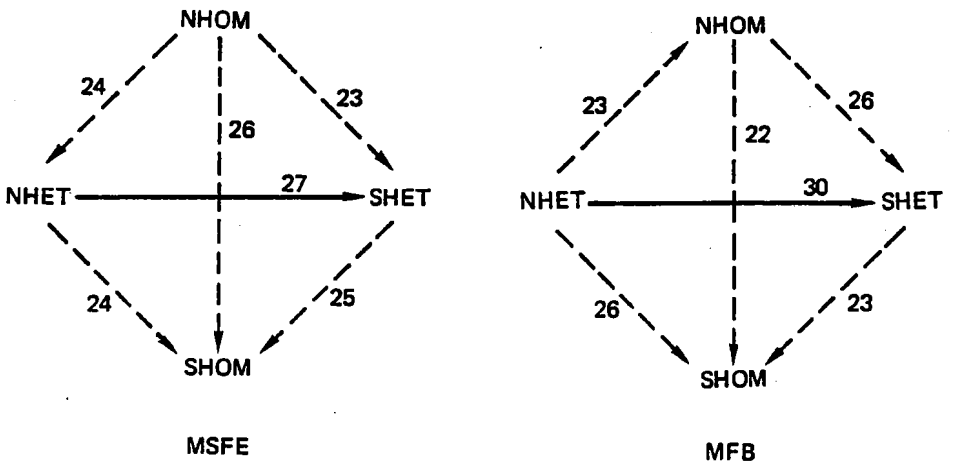
When the four-part model is considered, the split-sample analysis supports the choice of a smearing method (Fig. 5.3). SHOM is "significantly" better than NHOM in terms of MFB and weakly better in terms of MSFE. SHET is significantly better than NHET in terms of both MFB and MSFE. As in the case of the one-part and two-part models, the evidence for the choice of homoscedastic versus heteroscedastic methods is mixed. However, SHET dominates NHOM and SHOM weakly dominates NHET in terms of both MFB and MSFE. Hence, we should choose either SHOM or SHET, but cannot distinguish between them.

Three major conclusions come from this retransformation analysis. First, the analysis indicates the correct choice of either smearing or normal retransformation methods for each model. Second, within that choice, we are unable to distinguish between the homoscedastic and heteroscedastic versions. Third, the comparisons made earlier in this chapter have not been stacked against any particular model. In each case, we have been comparing the best version of each model.



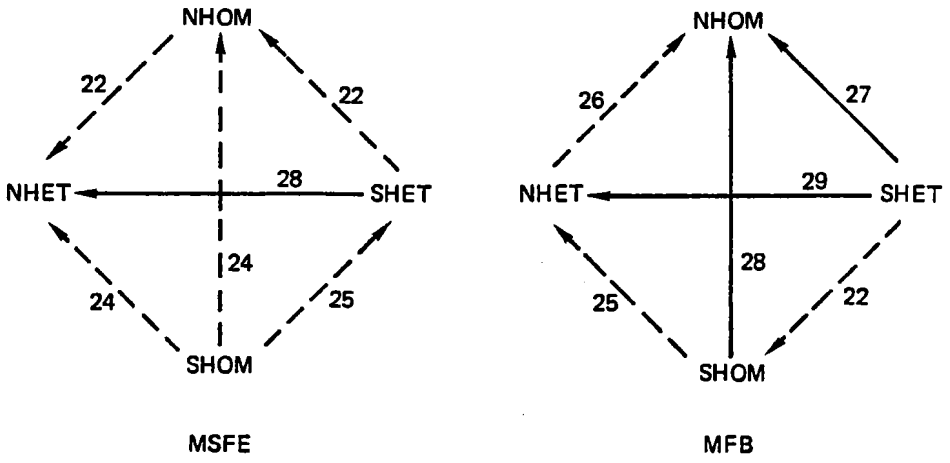
In each binary comparison in the diagram, the arrow points from the method with the higher number of site-year plans (with lower MSFE or MFB) to the one with the lower number. The number next to the arrow is the count for the former method. Solid lines are statistically significant at the 95-percent level; dashed lines are insignificant.

Fig. 5.1--One-part model: comparison of alternative retransformation methods



In each binary comparison in the diagram, the arrow points from the method with the higher number of site-year plans (with lower MSFE or MFB) to the one with the lower number. The number next to the arrow is the count for the former method. Solid lines are statistically significant at the 95-percent level; dashed lines are insignificant.

Fig. 5.2--Two-part model: comparison of alternative retransformation methods



In each binary comparison in the diagram, the arrow points from the method with the higher number of site-year plans (with lower MSFE or MFB) to the one with the lower number. The number next to the arrow is the count for the former method. Solid lines are statistically significant at the 95-percent level; dashed lines are insignificant.

Fig. 5.3--Four-part model: comparison of alternative retransformation methods

Chapter 6
CONCLUSION

The highly skewed distribution of medical expenditures, with a large number of nonspenders, makes reliable estimation and prediction difficult. Several alternative, plausible models provide different estimates of the response of demand for medical services to insurance coverage. Analyses of variance and covariance yield imprecise results, even for samples of over a thousand cases. The use of one-part and two-part transformed models improves the precision of the estimates and yields reasonably monotonic plan comparisons. However, the one-part model produces inconsistent results because it mishandles both the large number of nonspenders and the 10 percent of the sample with inpatient utilization. The two-part model corrects the former error but still produces inconsistent estimates because of the inpatient cases. The four-part model more accurately reflects the distribution of medical expenses than does either the one- or two-part model, and yields more precise estimates than ANOVA and ANOCOVA on untransformed expenditures.

Our split-sample analysis indicates that the two- and four-part models are more reliable for making forecasts than the other models. They are significantly better in terms of mean squared forecast error and weakly better in terms of mean forecast bias. However, we are unable to distinguish between the two- and four-part models. We prefer the four-part model over the two-part model because the latter is inconsistent. The split-sample analysis does not lead us to conclude that the four-part model is the better model in terms of forecasting to a future sample. But this should change as additional data become available. With larger data sets, the inconsistency in the two-part model will not decrease, but the degree of imprecision in the four-part model will.

The choice of the four-part model should not be considered final in any sense. Our experience has been that new data permit us to find problems that we could not see with smaller data sets, because

we can better evaluate the distributional assumptions in the analysis. In particular, more data will provide additional observations on the far-right tail of the distribution. Although relatively rare, these cases are very important because they contribute substantially to the overall average expense.

REFERENCES TO CHAPTERS 1 THROUGH 6

Balestra, P., and M. Nerlove, "Pooling Cross-Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas," *Econometrica*, Vol. 34, 1966, pp. 585-612.

Box, G. E. P., and D. R. Cox, "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 1962, pp. 211-245.

Brook, R. H., J. E. Ware, A. Davies-Avery, et al., "Overview of Adult Health Status Measures Fielded in Rand's Health Insurance Study," *Medical Care* (Supplement), Vol. 17, 1979, pp. 1-131.

Maddala, G. S., "The Use of Variance Components Models in Pooling Cross-Section and Time Series Data," *Econometrica*, Vol. 39, 1971, pp. 341-370.

Manning, W. G., C. N. Morris, J. P. Newhouse, et al., "A Two-Part Model of the Demand for Medical Care: Preliminary Results from the Health Insurance Study," in J. van der Gaag and M. Perlman (eds.), *Health, Economics, and Health Economics*, North Holland Publishing Co., Amsterdam, 1981.

McCarthy, Philip, "The Use of Balanced Half-Sample Replicate in Cross-Validation Studies," *Journal of the American Statistical Association*, Vol. 71, 1976, pp. 596-604.

Morris, C. N., "A Finite Selection Model for Experimental Design of the Health Insurance Study," in D. J. Aigner and C. N. Morris, "Experimental Design in Econometrics," *Journal of Econometrics*, Vol. 11, No. 1, September 1979, pp. 43-62.

Morris, C. N., J. P. Newhouse, and R. W. Archibald, "On the Theory and Practice of Obtaining Unbiased and Efficient Samples in Social Survey," in V. L. Smith (ed.), *Experimental Economics*, Vol. 1, JAI Press, Westport, Connecticut, 1979. Also in *Evaluation Studies Review Annual*, Vol. 5, Sage Publications, Beverly Hills, California, 1980.

Mosteller, Frederick, and J. W. Tukey, *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1977.

Mundlak, Y., "On the Pooling of Time Series and Cross-Section Data," *Econometrica*, Vol. 46, 1978, pp. 69-86.

Newhouse, J. P., "The Demand for Medical Care Services: A Retrospect and Prospect," in J. van der Gaag and M. Perlman (eds.), *Health, Economics, and Health Economics*, North Holland Publishing Co., Amsterdam, 1981.

Newhouse, J. P., "A Design for a Health Insurance Experiment," *Inquiry*, Vol. 11, 1974, pp. 5-27.

Newhouse, J. P., "Insurance Benefits, Out-of-Pocket Payments, and the Demand for Medical Care: A Review of the Literature," *Health Medical Services Review*, Vol. 1 (4), 1978, pp. 1, 3-15.

Newhouse, J. P., K. H. Marquis, C. N. Morris, et al., "Measurement Issues in the Second Generation of Social Experiments: The Health Insurance Study," in D. J. Aigner and C. N. Morris, "Experimental Design in Econometrics," *Journal of Econometrics*, Vol. 11, 1979, pp. 117-130.

Pregibon, Daryl, "Goodness-of-Link Tests for Generalized Linear Models," *Applied Statistics*, Vol. 29, No. 1, 1980, pp. 14-23.

Searle, S. R., *Linear Models*, John Wiley & Sons, Inc., New York, 1971.

Appendix A
TOBIT AND SELECTION MODEL ALTERNATIVES TO THE TWO-PART MODEL

The econometric literature has several models for data with continuous but limited dependent variables, such as medical expenses. Those models are the Tobit model proposed by Tobin (1958),[1] the two-part model discussed by Cragg (1971) and by Poirier and Ruud (1981), the selection model discussed by Heckman, and the adjusted Tobit model discussed by van de Ven and van Praag (1981). In this appendix, we describe each alternative model briefly and the reasons why we prefer the two-part model to the alternative models for analyzing the HIS data. This preference is specific to the HIS application and should not be construed to be a global pronouncement of the relative value of the alternative models.

THE TOBIT MODEL

The oldest of these models is the Tobit model. The limited dependent variable is modeled as a censored distribution. The model assumes that the unobserved uncensored error distribution is normal, whereas the observed error follows a censored normal distribution. More specifically,

$$
\begin{aligned}
&I_i = x_i\delta + \varepsilon_i, \quad \varepsilon \sim N(0,\sigma^2) , \\
&Y_i = I_i, \quad \text{if} \quad I_i \geq 0 , \qquad\qquad (A.1) \\
&Y_i = 0 , \quad \text{if} \quad I_i < 0 ,
\end{aligned}
$$

where Y_i is the dependent variable, I_i is the unobserved index based on individual characteristics, and x_i is a row vector.

---

[1] References are cited at the end of the appendix.

The expected expenditure is given by

$$E\ Y_i = \Phi(x_i\delta/\sigma) = x_i\delta + \sigma\phi(x_i\delta/\sigma) \ , \qquad (A.2)$$

which can be estimated by substituting appropriate estimates of δ and σ.

We rejected the Tobit model because the HIS data clearly do not fit the basic assumption of this model. As Fig. 3.5 indicates, the error distribution does not appear to be a censored normal distribution. The lower tail of the empirical distribution of positive expenses is quite smooth and approaches zero as the size approaches zero.

THE SELECTION MODEL

The second alternative to the two-part model is a two-equation self-selection model, as described by Heckman (1974, 1976, 1979).[2] The first equation specifies a probit censoring function that determines whether the positive expenditures are observed:

$$POSEXP_i = x_i\alpha_1 + \nu_{1i} \ . \qquad (A.3a)$$

If POSEXP is positive for an individual, then his expense EXPD is observed; otherwise 0 is observed. This equation is identical to Eq. (3.4a) in the two-part model. The second equation is an unconditional (uncensored) linear model of expenses:

$$EXPD_i = x_i\alpha_2 + \nu_{2i} \ , \qquad (A.3b)$$

where

[2] The econometric literature began with Amemiya (1973), Gronau (1974), and Heckman (1974).

$$\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \sim N(0,\Sigma) \ ,$$

and

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}, \qquad \sigma_{11} = 1 \ .$$

The expected expenditure is given by the unconditional (uncensored) expectation

$$E\ (EXPD_i) = x_i\alpha_2 \ ,$$

which can be estimated by substituting appropriate estimates of α_2.

If the errors in the two equations are correlated, then least squares on the *observed* positive expenditures will lead to biased estimates of α_2, the *unconditional* (uncensored) coefficient vector.

The self-selection model and the two-part model differ in their distributional assumptions. In the two-part model, we assume that positive expenses are (log) normally distributed.[3] The normal assumption is important only for retransforming the log dollars back to dollars. Any distribution would permit retransformations as long as E [exp (η_2)] is finite. In contrast, the self-selection model assumes that the error terms in (A.3a) and (A.3b) follow a bivariate normal distribution. It uses departures from normality and heteroscedasticity in the conditional distribution to estimate σ_{12}, σ_{22}, and α_2.

---

[3] In our discussion of the two-part model in Chapter 3, we assumed that some power transformation of nonzero expenses would yield a normal error distribution. The best power transformation turned out to be very close to the log.

We rejected the self-selection model for three reasons. The first reason is that it assumes that the functional form is known a priori and that the functional form must be such as to yield a bivariate normal error. Hence, the functional form cannot be estimated or evaluated from the data. The assumption in the self-selection model is untestable because the uncensored data are unobservable.

The second reason for not adopting the self-selection model is that its interpretation is inappropriate for modeling our expenditure data. The similarity between the self-selection problem and health expenditure data is not as real as it might appear. The self-selection model is an attempt to estimate the *unconditional* (uncensored) expenditure equation (A.3b), which describes the expenses that all individuals (including nonspenders) would have had if they were all spenders. However, for the study of health expenditure, we are not interested in this equation because we know that these individuals had zero expenses; unlike the self-selection problem, the zero spenders are not cases with missing expenses. Therefore, the conditional equation (3.4b), instead of (A.3b), is the equation of interest to us.

The third reason for rejecting the selection model is its poor numerical and statistical properties. The likelihood function of the selection model is known to have nonunique local maxima (Olsen, 1975). In contrast, the two-part model has a unique global maximum. Further, in the selection model, it is difficult to separate selection effects from heteroscedasticity and nonlinearity.[4] However, analytically we need to distinguish among these so that we can obtain

---

[4]The censored (positive) expenses are heteroscedastic and nonlinear in x. The variance of the positive expense is $\sigma_{22}[1 + \rho^2\lambda_i(z_i - \lambda_i)]$, where $z_i = -x_{2i}\alpha_2$ and λ_i is the reciprocal of the Mills' ratio. Equation A.6 indicates that the observed expenses are nonlinear in x.

a consistent estimate of the expected expense.[5] In contrast, in the two-part model, at least in Eq. (3.3b), we can use standard residual analyses to evaluate the appropriateness of our linear and additive specification.

THE ADJUSTED TOBIT MODEL

The third alternative model is the adjusted Tobit model, which specifies equations similar to (A.3a) and (A.3b), but calculates the expected expenditure by

$$E\ (\text{expense}) = \Pr\ (\text{expense} > 0) \cdot E\ (\text{expense} \mid \text{expense} > 0)\ . \tag{A.4}$$

This model is preferable to the self-selection model because the expected expenditure (A.4) refers to the *actual* expenditure instead of the expenditures that would be observed if everyone had positive expenditures.

The adjusted Tobit model and the two-part model are different specifications of the same problem. Both models specify a probit equation for the decision to have positive (nonzero) expenses. For log positive expense, the two-part model has a linear specification,

$$E_{TPM}\ (\log \text{expense} \mid \text{expense} > 0) = x\delta_2\ , \tag{A.5}$$

whereas the adjusted Tobit model (with a logarithmic scale) has

$$E_{ATM}\ (\log \text{expense} \mid \text{expense} > 0) = x\alpha_2 + \sigma_{12}\lambda(x\alpha_1)\ , \tag{A.6}$$

---

[5] Heteroscedasticity in σ_{22} has no effect on the expected value of α_2. However heteroscedasticity in the observed expenses may be interpreted as selection effects, which will alter the estimate of α_2. Similarly, significant coefficients for the reciprocal Mills' ratio have different interpretations, depending on whether they are selection effects or omitted nonlinearities in the underlying model.

where λ is the reciprocal of the Mills' ratio function.

We prefer the two-part model to the adjusted Tobit model because of its numerical simplicity and the absence of any strong theoretical justification for the additional complexity of the adjusted Tobit model. The adjusted Tobit model shares with the selection model the estimation and evaluation problems discussed at the end of the last section.

OMITTED VARIABLES

One misconception about the two-part model has been voiced by several colleagues on a preliminary version of this report. The misconception is that we *have to* assume independence between Eqs. (3.3a) and (3.3b). The misconception implies that the model contains no omitted variables that affect both the decision to seek care and the decision on the intensity of care.

Such an assumption of independence is *not* necessary for the two-part model. As we pointed out in our discussion of the two-part model in Chapter 3, the separability of the likelihood function of the two-part model is a result of the way the conditional densities are calculated. It does not depend on any independence assumption. The two equations can very well be correlated, say, because of omitted variables, and the separability will still hold.

REFERENCES TO APPENDIX A

Amemiya, T., "Regression Analysis When the Dependent Variable Is Truncated Normal," *Econometrica*, Vol. 41, 1973, pp. 997-1017.

Cragg, John G., "Some Statistical Models for Limited Dependent Variables with Applications to the Demand for Durable Goods," *Econometrica*, Vol. 39, 1971, pp. 829-844.

Gronau, R., "Wage Comparisons--A Selectivity Bias," *Journal of Political Economy*, Vol. 82, 1974, pp. 1119-1144.

Heckman, J., "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Sample Estimator for Such Models," *The Annals of Economic and Social Measurement*, Vol. 5, 1976, pp. 475-592.

Heckman, J., "Sample Bias as a Specification Error," *Econometrica*, Vol. 47, 1979, pp. 153-167.

Heckman, J., "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, Vol. 42, 1974, pp. 679-694.

Olsen, Randall, "The Analysis of Two-Variable Models When One of the Variables Is Dichotomous," Economics Department, Yale University, 1975 (unpublished manuscript).

Poirier, D. J., and P. A. Ruud, "On the Appropriateness of Endogenous Switching," *Journal of Econometrics*, Vol. 16, 1981, pp. 249-256.

Tobin, James, "Estimation of Relationships for Limited Dependent Variables, *Econometrica*, Vol. 26, 1958, pp. 24-36.

van de Ven, W. P. M. M., and B. M. S. van Praag, "Risk Aversion and Deductibles in Private Health Insurance: Applications of an Adjusted Tobit Model to Family Health Care Expenditures," in J. van der Gaag and M. Perlman (eds.), *Health, Economics, and Health Economics*, North Holland Publishing Co., Amsterdam, 1981.

Appendix B

SMEARING ESTIMATE: A NONPARAMETRIC RETRANSFORMATION METHOD

by *Naihua Duan*

B.1. INTRODUCTION

A monotonic transformation is often applied to observations recorded on a "natural" scale to achieve desirable statistical properties such as additivity, homoscedasticity, and normality. Certain analyses (e.g., fitting a least squares regression model) are carried out on the transformed scale, possibly combined with certain inferences, such as significance tests on comparisons of experimental treatments. However, it is very often desirable also to carry out certain procedures on the "natural" scale, e.g., prediction and forecasting. In doing so, one will be confronted with the problem of retransformation bias, namely, that unbiased and consistent quantities on the transformed scale usually do not retransform into unbiased or consistent quantities on the "natural" scale.

In this appendix, we propose a nonparametric method, the *smearing estimate*,[1] as an estimate of an individual's expected response on the "natural" scale. The essence of the procedure is to estimate the unknown error distribution by the empirical c.d.f. of the estimated regression residuals and then take the desired expectation with respect to the estimated error distribution. The estimate is weakly consistent under mild regularity conditions. In the case of logarithmic transformation, the efficiency of the smearing estimate relative to the corresponding parametric estimate, when the parametric model holds, is high over a wide range of parameter values.

In Section B.2, we present the retransformation problem and use an example to demonstrate the possible bias due to inappropriate use of the normal assumption. In Section B.3, we derive the smearing

---

[1]The terminology "smearing estimate" was originally coined by Professor Carl Morris for a related but different procedure.

estimate as an estimate of the "natural" scale expectation free from distributional assumptions on the error distribution F. The consistency property of the smearing estimate is established in Section B.4, and in Section B.5, we examine the efficiency of the smearing estimate compared with that of a parametric estimate when the parametric assumption is satisfied. In Section B.6, we discuss the prediction of group average. Our conclusions are briefly discussed in Section B.7.

B.2. THE RETRANSFORMATION PROBLEM

We denote the observations on the "natural" scale by Y_i, $i = 1, \ldots, n$, and the transformed observations by η_i, $i = 1, \ldots, n$, which are related by

$$\eta_i = g(Y_i), \quad Y_i = h(\eta_i), \quad h = g^{-1},$$

where g and h are assumed to be monotonic and continuously differentiable. To avoid the trivial cases, we also assume g and h to be nonlinear. We refer to g as the *transformation* and h as the *retransformation.*[2]

We consider a linear regression model on the transformed scale:

$$\eta_i = x_i\beta + \varepsilon_i ,$$

$$\varepsilon_i \sim F , \qquad \text{(i.i.d.)} ,$$

$$E\,\varepsilon_i = 0 ,$$

$$\text{Var}\,\varepsilon_i = \sigma^2 ,$$

where the x_i's are given row vectors of explanatory variables, β is a column vector of unknown parameters to be estimated, and the ε_i's

[2] We assume g and h to be known a priori.

are the residual errors. Although the error distribution F is usually assumed to be normal, we do not make this assumption. It is shown later in this section that inappropriate use of the normal assumption can lead to inconsistent prediction results.

For the assumed model, the minimum variance linear unbiased estimate of β is the least squares regression estimate on the transformed scale:

$$\hat{\beta} = (X'X)^{-1}X'\eta \ ,$$

where $X = (x_1', \ldots, x_n')'$ is the design matrix, assumed to have full rank, and $\eta = (\eta_1, \ldots, \eta_n)'$ is the transformed data vector. Moreover, for an individual with explanatory variables x_o, the prediction

$$x_o\hat{\beta} = x_o(X'X)^{-1}X'\eta$$

is the minimum variance linear unbiased estimate of the expectation of his response

$$E\ \eta_o = x_o\beta$$

on the transformed scale. Moreover, the regression coefficients $\hat{\beta}$, as well as the prediction $x_o\hat{\beta}$ for fixed x_o, are consistent if the design matrix is asymptotically nondegenerate.

In terms of the "natural" scale, it may seem "natural" to retransform the transformed scale prediction $x_o\hat{\beta}$ by $h = g^{-1}$ and use

$$h(x_o\hat{\beta})$$

to estimate the expectation of the individual's response

$$E\ Y_o = E\ h(\eta_o) = E\ h(x_o\beta + \varepsilon)$$

on the "natural" scale. However, the retransformed "natural" scale

prediction $h(x_o\hat{\beta})$ will no longer be unbiased, nor consistent, unless the transformation is linear, which we have assumed not to be the case. Actually, even if we know the true parameters, β, still $h(x_o\beta)$ is not the correct "estimate" of $E\ Y_o$:

$$E\ Y_o = E\ h(x_o\beta + \varepsilon) \neq h(x_o\beta) \ .$$

There is an extensive literature (e.g., Neyman and Scott, 1960)[3] devoted to the problem of estimating the "natural" scale expectation under the assumption that the error distribution is normal. We will refer to their results categorically as *normal theory estimates*.

It should be noted that the normality assumption plays very different roles in estimating the "natural" scale expectations and in estimating the regression coefficients. For estimating the regression coefficients, whether the true error distribution is normal or not, the least squares estimate, which is the maximum likelihood estimate under the normal assumption, is consistent and minimum variance linear unbiased. When the true error distribution is not normal, the normality assumption will only affect the efficiency of our estimate (Cox and Hinkley, 1968). If we know the form of the true error distribution, we can sometimes derive alternative estimates that will be more efficient than the least squares estimate. However, for estimating the "natural" scale expectation, an incorrect normality assumption can lead to inconsistent estimates.

For example, in the case of a logarithmic transformation with normally distributed error, the "natural" scale expectation is

$$\exp\ (x_o\beta + \sigma^2/2) \ ,$$

where

$$\sigma^2 = \text{Var}\ \varepsilon \ .$$

---

[3]References are cited at the end of the appendix.

The expectation can be estimated consistently by the normal theory estimate

$$\exp\ (x_o\hat{\beta} + \hat{\sigma}^2/2)\ ,$$

where $\hat{\beta}$ denotes the least squares regression coefficients and $\hat{\sigma}^2$ denotes the mean squared error.

Whether the true error distribution is normal or not, the above estimate is consistent for $\exp\ (x_o\beta + \sigma^2/2)$. However, it might not be consistent for $E\ Y_o$. For example, if the true error distribution is actually a mixture of two normal distributions,

$$\left.\begin{array}{ll} \varepsilon \sim N(0,\ .95\ \sigma^2) & \text{with probability } .995 \\ \sim N(0,\ 10.95\ \sigma^2) & \text{with probability } .005 \end{array}\right\} \quad (\text{Var } \varepsilon = \sigma^2)\ ,$$

then the "natural" scale expectation is

$$E\ Y_o = .995\ \exp\ (x_o\beta + .475\ \sigma^2)$$

$$+ .005\ \exp\ (x_o\beta + 5.475\ \sigma^2)\ .$$

For $\sigma^2 = 1$, we have

$$E\ Y_o = 2.79\ \exp\ (x_o\beta)\ .$$

The normal theory estimate will converge almost surely to

$$\exp\ (x_o\beta + \sigma^2/2) = 1.65\ \exp\ (x_o\beta)\ ,$$

which has a 41-percent asymptotic bias.

B.3. THE SMEARING ESTIMATE

Our goal is to estimate the "natural" scale expectation

$$E\ Y_o = E\ h(x_o\beta + \varepsilon) = \int h(x_o\beta + \varepsilon)\ dF(\varepsilon)\ .$$

Without knowing the error distribution function F or a reliable parametric form for it, we will estimate F by the empirical c.d.f. of the estimated residuals

$$\hat{F}_n(e) = \frac{1}{n} \sum_{i=1}^{n} I\{\hat{\varepsilon}_i \leq e\}\ ,$$

where $\hat{\varepsilon}_i = \eta_i - x_i\hat{\beta}$ denotes the least squares residual and $I\{\cdot\}$ denotes the indicator function of the event "$\cdot$". Duan (1980) proved under mild regularity conditions that the nonparametric estimate $\hat{F}_n$ is strongly consistent in the uniform norm:

$$\sup_e |\hat{F}_n(e) - F(e)| \to 0 \qquad \text{(a.s.)}\ .$$

As is usual in nonparametric analyses, a population quantity with an expression in terms of the true c.d.f. can be estimated by the corresponding expression in terms of the empirical c.d.f. For example, the population mean

$$\mu = \int x\ dF(x)$$

can be estimated nonparametrically by the sample mean

$$\bar{x} = \int x\ d\hat{F}_n(x)\ .$$

Similarly, we estimate $E\ Y_o$ by substituting the unknown c.d.f. F by its empirical estimate $\hat{F}_n$:

$$\tilde{E}\ Y_o = \int h(x_o\beta + \varepsilon)\ d\hat{F}_n(\varepsilon) = \frac{1}{n} \sum_{i=1}^{n} h(x_o\beta + \hat{\varepsilon}_i)\ .$$

Further substituting the regression parameters β by their least squares estimates $\hat{\beta}$, we have the estimate

$$\hat{E}\, Y_o = \int h(x_o\hat{\beta} + \varepsilon)\, d\hat{F}_n(\varepsilon) = \frac{1}{n}\sum_{i=1}^{n} h(x_o\hat{\beta} + \hat{\varepsilon}_i) , \qquad (B.3.1)$$

which will be referred to as the *smearing estimate*.

As will be shown in Section B.4, the smearing estimate derived in this way is consistent under mild regularity assumptions on the error distribution F and the design matrix X.

B.4. CONSISTENCY OF THE SMEARING ESTIMATE

Assuming that h is continuously differentiable, we take the first-order Taylor's expansion:

$$h(x_o\hat{\beta} + \hat{\varepsilon}_i) = h(x_o\beta + \varepsilon_i) + \delta_i \cdot h'(x_o\beta + \varepsilon_i + \theta_i\delta_i) ,$$

where

$$0 \leq \theta_i \leq 1$$

$$\delta_i = (x_o\hat{\beta} + \hat{\varepsilon}_i) - (x_o\beta + \varepsilon_i) = (x_o - x_i)(X'X)^{-1}X'\varepsilon .$$

The smearing estimate can be decomposed as follows:

$$\hat{E}\, Y_o = \frac{1}{n}\sum_{i=1}^{n} h(x_o\hat{\beta} + \hat{\varepsilon}_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} h(x_o\beta + \varepsilon_i) + \frac{1}{n}\sum_{i=1}^{n} \delta_i \cdot h'(x_o\beta + \varepsilon_i + \theta_i\delta_i) . \qquad (B.4.1)$$

By the strong law of large numbers, the first term on the right-hand side of (B.4.1) is strongly consistent for the "natural" scale expectation $E\, Y_o$. It remains to show that the second term is stochastically small in some sense.

It follows from the Cauchy-Schwarz inequality that the square of the second term in (B.4.1) is bounded from above by the product

$$\frac{1}{n} \sum_{i=1}^{n} \delta_i^2 \cdot \frac{1}{n} \sum_{i=1}^{n} [h'(x_o\beta + \varepsilon_i + \theta_i\delta_i)]^2 \ . \tag{B.4.2}$$

The proof of the following lemma is given in Addendum I.

Lemma 1. Assume that (1) the retransformation h is continuously differentiable, (2) X contains the intercept, and (3) $X'X/n \to \ddagger$ positive definite; then

$$\sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} [(x_o - x_i)(X'X)^{-1}X'\varepsilon]^2 = O_p(1) \ . \quad \|$$

It follows immediately from Lemma 1 that the first factor in (B.4.2) has order $O_p(1/n)$; in particular, it converges to zero in probability.

It also follows from Lemma 1 that we can choose M large enough such that, for n large enough, the inequality

$$\sum_{i=1}^{n} \delta_i^2 < M^2 \tag{B.4.3}$$

will hold with probability arbitrarily close to one. When (B.4.3) holds, we have

$$|\delta_i| \leq M \ , \qquad i = 1, \ldots, n \ ,$$

$$|h'(x_o\beta + \varepsilon_i + \theta_i\delta_i)| \leq \sup_{|t|\leq M} |h'(x_o\beta + \varepsilon_i + t)| \ ;$$

thus

$$\frac{1}{n}\sum_{i=1}^{n} [h'(x_o\beta + \varepsilon_i + \theta_i\delta_i)]^2 \leq \frac{1}{n}\sum_{i=1}^{n} \sup_{|t|\leq M} [h'(x_o\beta + \varepsilon_i + t)]^2 \ . \tag{B.4.4}$$

By the strong law of large numbers, the right-hand side of (B.4.4) converges almost surely to

$$E \sup_{|t|\leq M} [h'(x_o\beta + \varepsilon + t)]^2 \tag{B.4.5}$$

if the expectation is finite.

To summarize, if the expectation (B.4.5) is finite for all $M > 0$, the second factor in (B.4.2) is bounded from above, with probability arbitrarily close to one, by a sequence of random variables that converge almost surely to a finite constant. In other words, the second factor in (B.4.2) is stochastically bounded. Thus, we have proved

Theorem 2. Assume that (1) the retransformation h is continuously differentiable, (2) X contains the intercept, (3) $X'X/n \to \ddagger$ positive definite, and (4) the expectation (B.4.5) is finite for all $M > 0$; then the smearing estimate (B.3.1) is weakly consistent. ||

For most popular transformations, the supremum in (B.4.5) can be evaluated at the end points. For example, if $|h'|$ is monotonic, we have

$$E \sup_{|t|\leq M} [h'(x_o\beta + \varepsilon + t)]^2 \leq E\,[h'(x_o\beta + \varepsilon + M)]^2 + E\,[h'(x_o\beta + \varepsilon - M)]^2 \ .$$

The moment condition (4) in Theorem 2 can be replaced by

$$(4') \qquad E\,[h'(c + \varepsilon)]^2 < +\infty \quad \text{for all } c\ ,$$

which is usually easy to check under hypothesized true error distribution. For example, for power transformations

$$\eta = g(Y) = Y^{\alpha}\ , \qquad \alpha \neq 0\ ,$$

$$Y = h(\eta) = \eta^{1/\alpha}\ ,$$

the desired moment condition is that

$$E\,(c + \varepsilon)^{2[(1/\alpha)-1]} < +\infty$$

for all c, which will be satisfied for normal error distribution if $0 < \alpha < 1$. For the logarithmic transformation

$$\eta = \log\,(Y)\ , \qquad Y = \exp\,(\eta)\ .$$

The desired moment condition reduces to

$$E\,e^{2\varepsilon} < +\infty\ ,$$

which is satisfied for the normal error distribution.

B.5. EFFICIENCY OF THE SMEARING ESTIMATE

If the error distribution is indeed normal, the normal theory estimate and the smearing estimate are both consistent, but the normal theory estimate can be more efficient. In this section we examine the loss of efficiency of the smearing estimate relative to the normal theory estimate when the error distribution is indeed normal.

For simplicity, we consider only the logarithmic transformation in this section. The normal theory estimate (NE) and the smearing estimate (SE) are

$$NE = e^{x_o\hat{\beta}+\hat{\sigma}^2/2}$$

$$SE = e^{x_o\hat{\beta}} \cdot \frac{1}{n} \sum_{i=1}^{n} e^{\hat{\varepsilon}_i} ,$$

where $\hat{\beta}$ are the least squares regression coefficients, $\hat{\sigma}^2$ is the residual variance, and $\hat{\varepsilon}_i$ is the least squares residual. It follows from the normal assumption that the regression coefficients $\hat{\beta}$ are stochastically independent of the residuals $\hat{\varepsilon}$; in particular, we have

$$e^{x_o\hat{\beta}} \text{ independent of } e^{\hat{\sigma}^2/2} ,$$

$$e^{x_o\hat{\beta}} \text{ independent of } \frac{1}{n} \sum_{i=1}^{n} e^{\hat{\varepsilon}_i} .$$

Therefore, we can evaluate the variances as follows:

$$\text{Var (NE)} = E^2 (e^{x_o\hat{\beta}}) \cdot \text{Var} (e^{\hat{\sigma}^2/2}) + \text{Var} (e^{x_o\hat{\beta}}) \cdot E^2 (e^{\hat{\sigma}^2/2}) + \text{Var} (e^{x_o\hat{\beta}}) \cdot \text{Var} (e^{\hat{\sigma}^2/2}) , \quad \text{(B.5.1)}$$

$$\text{Var (SE)} = E^2 (e^{x_o\hat{\beta}}) \cdot \text{Var} (\frac{1}{n} \sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) + \text{Var} (e^{x_o\hat{\beta}}) \cdot E^2 (\frac{1}{n} \sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) + \text{Var} (e^{x_o\hat{\beta}}) \cdot \text{Var} (\frac{1}{n} \sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) . \quad \text{(B.5.2)}$$

The proof of the following lemma is given in Addendum II.

Lemma 3. Assume that (1) $X'X/n \to \Sigma$ positive definite, (2) X contains the intercept, and (3) $\varepsilon \sim N(0,\sigma^2)$; then

$$E\,(e^{x_o\hat{\beta}}) \to e^{x_o\beta} ,$$

$$n\ \mathrm{Var}\,(e^{x_o\hat{\beta}}) \to (x_o\Sigma^{-1}x_o') \cdot \sigma^2 e^{2x_o\beta} ,$$

$$E\,(e^{\hat{\sigma}^2/2}) \to e^{\sigma^2/2} ,$$

$$n\ \mathrm{Var}\,(e^{\hat{\sigma}^2/2}) \to \tfrac{1}{2}\sigma^4 e^{\sigma^2} ,$$

$$E\,(\frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) \to e^{\sigma^2/2} ,$$

$$n\ \mathrm{Var}\,(\frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) \to e^{\sigma^2}(e^{\sigma^2} - 1 - \sigma^2) . \qquad \|$$

Substituting the limits in Lemma 3 into the previous variance formulae, (B.5.1) and (B.5.2), we have

Theorem 4. Assume that (1) $X'X/n \to \Sigma$ positive definite, (2) X contains the intercept, and (3) $\varepsilon \sim N(0,\sigma^2)$; then

$$n\ \mathrm{Var}\,(NE) \to [(x_o\Sigma^{-1}x_o')\sigma^2 + \tfrac{1}{2}\sigma^4]e^{2x_o\beta+\sigma^2} ,$$

$$n\ \mathrm{Var}\,(SE) \to [(x_o\Sigma^{-1}x_o')\sigma^2 + (e^{\sigma^2} - 1 - \sigma^2)]e^{2x_o\beta+\sigma^2} .$$

The relative efficiency of the smearing estimate to the normal theory estimate is

$$\text{rel. eff.} = \frac{\text{Var (NE)}}{\text{Var (SE)}} = \frac{(x_o \ddagger^{-1} x_o')\sigma^2 + \tfrac{1}{2}\sigma^4}{(x_o \ddagger^{-1} x_o')\sigma^2 + (e^{\sigma^2} - 1 - \sigma^2)} \; . \qquad \| \tag{B.5.3}$$

(Note that $\frac{1}{2}\sigma^4$ is the leading term in the Taylor's series expansion of $e^{\sigma^2} - 1 - \sigma^2$.)

The relative efficiency depends on both σ^2 and $x_o \ddagger^{-1} x_o'$. If x_o is sampled randomly from the same population as x_i's, we have

$$E\, x_o \ddagger^{-1} x_o' = \text{tr} \ddagger^{-1} E\, x_o' x_o$$

$$= \text{tr} \ddagger^{-1} \ddagger \qquad (\ddagger = E\, x'x)$$

$$= \text{rank}(X) \; ;$$

thus $x_o \ddagger^{-1} x_o'$ is of the same order as rank(X). Table B.1 contains the relative efficiency for several values of σ^2 and $x_o \ddagger^{-1} x_o'$. For σ^2 near or less than 1.5, the relative efficiency is quite high. For large σ^2, the relative efficiency drops drastically. Under the assumed model, the "natural" scale responses follow a lognormal distribution, with σ^2 being the shape parameter: large σ^2 indicates large skewness.

For large σ^2, while the normal theory estimate is more efficient than the smearing estimate when the normal assumption is true, it can also be more sensitive to departures from normality. As an illustration, we will consider again the example used in the introduction:

$$\left.\begin{array}{ll} \varepsilon \sim N(0, .95\sigma^2) & \text{with probability } .995 \\ \sim N(0, 10.95\sigma^2) & \text{with probability } .005 \end{array}\right\} \quad (\text{Var } \varepsilon = \sigma^2) \; .$$

Table B.2 provides the relative bias of the normal theory estimate. The relative bias increases with σ^2; for σ^2 near or larger

Table B.1

RELATIVE EFFICIENCY OF THE SMEARING ESTIMATE TO THE NORMAL THEORY ESTIMATE WHEN THE NORMALITY ASSUMPTION IS SATISFIED

	$x_o \ddagger^{-1} x_o'$ ($\approx$ rank X)					
σ^2	1	2	3	5	10	20
0.50	0.96	0.98	0.99	0.99	1.00	1.00
0.75	0.92	0.95	0.97	0.98	0.99	0.99
1.00	0.87	0.92	0.94	0.96	0.98	0.99
1.50	0.75	0.83	0.87	0.91	0.95	0.97
2.00	0.63	0.72	0.77	0.83	0.90	0.95
3.00	0.39	0.48	0.54	0.63	0.75	0.85
5.00	0.12	0.15	0.17	0.22	0.32	0.46
10.00	0.00	0.00	0.00	0.00	0.01	0.01

Table B.2

RELATIVE BIAS OF THE NORMAL THEORY ESTIMATE UNDER THE MIXTURE MODEL

σ^2	Relative Bias[a] (%)
.50	-4
.75	-16
1.00	-41
1.50	-90
2.00	-99
3.00	-100

[a]Under the assumed model, the relative bias is

$$\frac{E\,(NE) - E\,Y_o}{E\,Y_o}$$

$$= \frac{e^{x_o\beta + .5\sigma^2}}{e^{x_o\beta}(.995e^{.475\sigma^2} + .005e^{4.975\sigma^2})} - 1 .$$

than one, the normal theory estimate is severely biased.

B.6. PREDICTION FOR GROUP AVERAGE

Up until now, our discussion has been restricted to the prediction of one fixed individual with explanatory variables x_o. One might also be interested in the prediction for the average of a group of individuals, with explanatory variables ξ_j, $j = 1, \ldots, J$. For simplicity we will consider J and ξ_j's as fixed. We have the estimates

$$SE^* = \frac{1}{J} \sum_{j=1}^{J} e^{\xi_j \hat{\beta}} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} e^{\hat{\varepsilon}_i}\right) ,$$

$$NE^* = \frac{1}{J} \sum_{j=1}^{J} e^{\xi_j \hat{\beta}} \cdot e^{\hat{\sigma}^2/2} .$$

The consistency of the smearing estimate for the group average follows immediately from the consistency for individual predictions. The proof of the following lemma is given in Addendum III.

Lemma 5. Assume that (1) $X'X/n \to \Sigma$ positive definite, (2) X contains the intercept, and (3) $\varepsilon \sim N(0,\sigma^2)$; then

$$E\left(\frac{1}{J} \sum_{j=1}^{J} e^{\xi_j \hat{\beta}}\right) \to \frac{1}{J} \sum_{j=1}^{J} e^{\xi_j \beta}$$

$$n \operatorname{Var}\left(\frac{1}{J} \sum_{j=1}^{J} e^{\xi_j \hat{\beta}}\right) \to \sigma^2 \cdot \left(\frac{1}{J} \sum_{j=1}^{J} e^{\xi_j \beta} \xi_j\right) \Sigma^{-1} \left(\frac{1}{J} \sum_{j=1}^{J} e^{\xi_j \beta} \xi_j\right)' . \quad \|$$

Substituting the limits in Lemmas 3 and 5 into the appropriate variance formulae, we have the following theorem:

Theorem 6. Under the same assumptions as in Lemma 3,

$$n \text{ Var } (NE^*) \to \left(\frac{1}{J} \sum_j e^{\xi_j \beta}\right)^2 \cdot \tfrac{1}{2}\sigma^4 e^{\sigma^2}$$

$$+ \left(\frac{1}{J} \sum_j e^{\xi_j \beta} \xi_j\right) \Psi^{-1} \left(\frac{1}{J} \sum_j e^{\xi_j \beta} \xi_j\right)' \cdot \sigma^2 e^{\sigma^2} ,$$

$$n \text{ Var } (SE^*) \to \left(\frac{1}{J} \sum_j e^{\xi_j \beta}\right)^2 \cdot (e^{\sigma^2} - 1 - \sigma^2) e^{\sigma^2}$$

$$+ \left(\frac{1}{J} \sum_j e^{\xi_j \beta} \xi_j\right) \Psi^{-1} \left(\frac{1}{J} \sum_j e^{\xi_j \beta} \xi_j\right)' \cdot \sigma^2 e^{\sigma^2} .$$

The relative efficiency is

$$\text{rel. eff.} = \frac{\text{Var } (NE^*)}{\text{Var } (SE^*)} = \frac{\omega\sigma^2 + \tfrac{1}{2}\sigma^4}{\omega\sigma^2 + (e^{\sigma^2} - 1 - \sigma^2)} ,$$

where

$$\omega = \left(\sum_j \frac{e^{\xi_j \beta}}{\sum_j e^{\xi_j \beta}} \xi_j \right) \Psi^{-1} \left(\sum_j \frac{e^{\xi_j \beta}}{\sum_j e^{\xi_j \beta}} \xi_j \right)' . \qquad \|$$

If we further assume that the ξ_j's are sampled from a normal distribution with mean ξ and covariance matrix T,[4] we have the following lemma, which is proved in Addendum IV.

Lemma 7. Assume that $\xi \sim N(\bar{\xi}, T)$; we have the almost sure convergence for ω as $J \to +\infty$:

---

[4] $\Psi = \xi\xi' + T$.

$$\omega \to \left(\frac{E\, e^{\xi\beta}\xi}{E\, e^{\xi\beta}}\right)\Sigma^{-1}\left(\frac{E\, e^{\xi\beta}\xi}{E\, e^{\xi\beta}}\right)' = (\bar{\xi} + \beta'T)\Sigma^{-1}(\bar{\xi} + \beta'T)' \ . \qquad \| \quad \text{(a.s.)}$$

Corollary 8. Under the assumptions in Lemmas 3 and 7, the relative efficiency is

$$\text{rel. eff.} = \frac{\text{Var } (NE^*)}{\text{Var } (SE^*)} = \frac{\omega^*\sigma^2 + \frac{1}{2}\sigma^4}{\omega^*\sigma^2 + (e^{\sigma^2} - 1 - \sigma^2)} \ ,$$

where

$$\omega^* = (\bar{\xi} + \beta'T)\Sigma^{-1}(\bar{\xi} + \beta'T)' \ . \qquad \|$$

B.7. CONCLUSION

In this appendix, we propose the smearing estimate as a nonparametric estimate of expected response on the untransformed scale. The estimate is consistent under mild regularity conditions and usually attains high efficiency relative to parametric estimates. It can be viewed as a low-premium insurance policy against departures from parametric distributional assumptions.

Addendum I
PROOF OF LEMMA 1

The sum in the lemma can be expressed as follows:

$$\sum_{i=1}^{n} [(x_o - x_i)(X'X)^{-1}X'\varepsilon]^2 = [\sqrt{n}\, x_o(X'X)^{-1}X'\varepsilon]^2$$

$$- 2[\sqrt{n}\, x_o(X'X)^{-1}X'\varepsilon]$$

$$\cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i(X'X)^{-1}X'\varepsilon$$

$$+ \sum_{i=1}^{n} [x_i(X'X)^{-1}X'\varepsilon]^2 \quad . \qquad \text{(B.4.3)}$$

We prove in three segments that the sum is asymptotically bounded.

(i) $$E\,[\sqrt{n}\, x_o(X'X)^{-1}X'\varepsilon] = 0 \ ,$$

$$\text{Var}\,[\sqrt{n}\, x_o(X'X)^{-1}X'\varepsilon] = \sigma^2 x_o(X'X/n)^{-1}x_o' \ .$$

Assuming that $X'X/n$ converges to a positive definite matrix $\ddagger$,[5] we have

$$\text{Var}\,[\sqrt{n}\, x_o(X'X)^{-1}X'\varepsilon] \to \sigma^2 x_o \ddagger^{-1} x_o' \ .$$

---

[5] The assumption is much stronger than we need; what really needs to be asserted is that $x_o(X'X/n)^{-1}x_o'$ will remain bounded. Nevertheless, the present assumption will be satisfied for many important problems, such as when the covariate x_i's are sampled randomly from a fixed parent population.

By the Chebyschev inequality, we have

$$P\left\{\left|\sqrt{n}\, x_o(X'X)^{-1}X'\varepsilon\right| > c\right\} \leq \frac{\text{Var}\,[\sqrt{n}\, x_o(X'X)^{-1}X'\varepsilon]}{c^2} \to \frac{\sigma^2 x_o(X'X)^{-1}x_o'}{c^2}$$

for any $c > 0$. For c large enough, we can make the limit arbitrarily small. Therefore

$$\sqrt{n}\, x_o(X'X)^{-1}X'\varepsilon = O_p(1) .$$

$$\text{(ii)}\quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i(X'X)^{-1}X'\varepsilon = \frac{1}{\sqrt{n}}\, 1'X(X'X)^{-1}X'\varepsilon$$

$$= \frac{1}{\sqrt{n}}\, 1'\varepsilon \quad \text{[assume X contains the intercept]}$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i$$

$$= O_p(1) \quad \text{[converges in law to } N(0,\sigma^2)\text{]} .$$

$$\text{(iii)}\quad E \sum_{i=1}^{n} [x_i(X'X)^{-1}X'\varepsilon]^2 = \sum_{i=1}^{n} E\, x_i(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}x_i'$$

$$= \sigma^2 \sum_{i=1}^{n} x_i(X'X)^{-1}x_i' \quad [E\,\varepsilon\varepsilon' = \sigma^2 I]$$

$$= \sigma^2\, \text{tr}\, X(X'X)^{-1}X'$$

$$= k\sigma^2 . \qquad [k = \text{rank } X] .$$

By the Markov inequality, we have

$$P\left\{\sum_{i=1}^{n}[x_i(X'X)^{-1}X'\varepsilon]^2 > c\right\} \leq \frac{k\sigma^2}{c} .$$

The limit can be made arbitrarily small by taking c large enough. Therefore

$$\sum_{i=1}^{n}[x_i(X'X)^{-1}X'\varepsilon]^2 = o_p(1) .$$

Addendum II

PROOF OF LEMMA 3

(i) $\quad E\,(e^{x_o\hat{\beta}}) \to e^{x_o\beta}, \qquad n\,\mathrm{Var}\,(e^{x_o\hat{\beta}}) \to x_o \mathfrak{I}^{-1} x_o' \cdot \sigma^2 e^{2x_o\beta}$.

The least squares estimate $\hat{\beta}$ is distributed as

$$\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1}) \ .$$

Therefore,

$$x_o\hat{\beta} \sim N(x_o\beta, \sigma^2 x_o (X'X)^{-1} x_o') \ ,$$

$$E\,(e^{x_o\hat{\beta}}) = e^{x_o\beta + \frac{1}{2}\sigma^2 x_o (X'X)^{-1} x_o'}$$

$$= e^{x_o\beta + (1/2n)\sigma^2 x_o (X'X/n)^{-1} x_o'}$$

$$\to e^{x_o\beta} \qquad [X'X/n \to \mathfrak{I},\ (1/2n)\sigma^2 x_o (X'X/n)^{-1} x_o' \to 0] \ ;$$

$$\mathrm{Var}\,(e^{x_o\hat{\beta}}) = e^{2x_o\beta + \sigma^2 x_o (X'X)^{-1} x_o'} (e^{\sigma^2 x_o (X'X)^{-1} x_o'} - 1) \ ,$$

$$e^{2x_o\beta + \sigma^2 x_o (X'X)^{-1} x_o'} = e^{2x_o\beta + (1/n)\sigma^2 x_o (X'X/n)^{-1} x_o'}$$

$$\to e^{2x_o\beta} \ ;$$

$$n(e^{\sigma^2 x_o (X'X)^{-1} x_o'} - 1) = n(e^{(1/n)\sigma^2 x_o (X'X/n)^{-1} x_o'} - 1)$$

$$\to \sigma^2 x_o \mathfrak{I}^{-1} x_o' \ .$$

Thus

$$n \text{ Var } (e^{x_o\hat{\beta}}) \to x_o \ddagger^{-1} x_o' \sigma^2 e^{2x_o\beta} . \quad \|$$

(ii) $$E\,(e^{\hat{\sigma}^2/2}) \to e^{\sigma^2/2} , \qquad n \text{ Var } (e^{\hat{\sigma}^2/2}) \to \tfrac{1}{2}\sigma^4 e^{\sigma^2} .$$

The least squares mean squared error is distributed as

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-k} \chi^2_{n-k} ,$$

where k = rank X, n - k = d.f. of residual sum of squares. It follows from Gamma integration that[6]

$$E\,(e^{\hat{\sigma}^2/2}) = \left(\frac{n-k}{n-k-\sigma^2}\right)^{(n-k)/2}$$

$$= \frac{1}{\left(1 - \frac{\sigma^2/2}{(n-k)/2}\right)^{(n-k)/2}}$$

$$\to \frac{1}{e^{-\sigma^2/2}} = e^{\sigma^2/2} \quad \text{as} \quad n \to \infty \qquad (k \text{ fixed}).$$

It also follows from Gamma integration that[7]

$$\text{Var } (e^{\hat{\sigma}^2/2}) = \left(\frac{n-k}{n-k-2\sigma^2}\right)^{(n-k)/2} - \left(\frac{n-k}{n-k-\sigma^2}\right)^{n-k}$$

---

[6]The expectation exists and equals the right-hand side of the formula only if $n - k < \sigma^2$, which for any fixed σ^2 will hold for n large enough.

[7]The variance exists and equals the right-hand side of the formula only if $n - k > 2\sigma^2$, which for any fixed σ^2 will also hold for n large enough.

$$= \frac{1}{\left(1 - \frac{\sigma^2}{(n-k)/2}\right)^{(n-k)/2}} - \frac{1}{\left(1 - \frac{\sigma^2}{n-k}\right)^{n-k}}$$

$$= \frac{\left(1 - \frac{\sigma^2}{n-k}\right)^{n-k} - \left(1 - \frac{\sigma^2}{(n-k)/2}\right)^{(n-k)/2}}{\left(1 - \frac{\sigma^2}{(n-k)/2}\right)^{(n-k)/2}\left(1 - \frac{\sigma^2}{n-k}\right)^{n-k}} .$$

The denominator converges to

$$e^{-\sigma^2} \cdot e^{-\sigma^2} = e^{-2\sigma^2} .$$

For the numerator, we will replace $n - k$ by $1/r$, and take Taylor's expansion around $r = 0$ $(n = +\infty)$:

$$\left(1 - \frac{\sigma^2}{n-k}\right)^{n-k} - \left(1 - \frac{\sigma^2}{(n-k)/2}\right)^{(n-k)/2}$$

$$= (1 - r\sigma^2)^{1/r} - (1 - 2r\sigma^2)^{1/2r} .$$

Denoting the above expression by $f(r)$, we have

$$\lim_{r \to 0} f(r) = e^{-\sigma^2} - e^{-\sigma^2} = 0 .$$

Differentiating with respect to r, we have

$$f'(r) = (1 - r\sigma^2)^{1/r} \cdot \left[-\frac{1}{r^2} \log (1 - r\sigma^2) + \frac{1}{r} \cdot \frac{-\sigma^2}{1 - r\sigma^2}\right]$$

$$- (1 - 2r\sigma^2)^{1/2r} \cdot \left[-\frac{1}{2r^2} \log (1 - 2r\sigma^2) + \frac{1}{2r} \cdot \frac{-2\sigma^2}{1 - 2r\sigma^2}\right] .$$

Note that

$$-\frac{1}{r^2}\log(1-r\sigma^2)+\frac{1}{r}\cdot\frac{-\sigma^2}{1-r\sigma^2}$$

$$=-\frac{1}{r^2(1-r\sigma^2)}\cdot[(1-r\sigma^2)\log(1-r\sigma^2)+r\sigma^2]$$

$$=-\frac{1}{r^2(1-r\sigma^2)}\cdot[(1-r\sigma^2)(-r\sigma^2-\frac{r^2\sigma^4}{2}+o(r^2))+r\sigma^2]$$

$$=-\frac{1}{r^2(1-r\sigma^2)}\cdot[-r\sigma^2-\frac{r^2\sigma^4}{2}+r^2\sigma^4+o(r^2)+r\sigma^2]$$

$$=-\frac{1}{r^2(1-r\sigma^2)}\cdot[\tfrac{1}{2}r^2\sigma^4+o(r^2)]$$

$$=-\frac{\sigma^4}{2(1-r\sigma^2)}+o(1)$$

$$=-\tfrac{1}{2}\sigma^4+o(1)\ ,$$

where $o(r^\alpha)$ denotes a quantity that satisfies

$$\frac{o(r^\alpha)}{r^\alpha}\to 0 \quad \text{as} \quad r\to 0 \qquad (n\to+\infty)\ .$$

Similarly, we have

$$-\frac{1}{2r^2}\log(1-2r\sigma^2)+\frac{1}{2r}\cdot\frac{-2\sigma^2}{(1-2r\sigma^2)}$$

$$=-\frac{1}{2r^2(1-2r\sigma^2)}\cdot[(1-2r\sigma^2)\log(1-2r\sigma^2)+2r\sigma^2]$$

$$= -\frac{1}{2r^2(1 - 2r\sigma^2)} \cdot [\tfrac{1}{2}(2r\sigma^2)^2 + o(r^2)]$$

$$= -\sigma^4 + o(1) .$$

Therefore,

$$f'(r) \to e^{-\sigma^2} \cdot (-\frac{\sigma^4}{2}) - e^{-\sigma^2} \cdot (-\sigma^4) = \tfrac{1}{2}\sigma^4 e^{-\sigma^2} \quad \text{as} \quad r \to 0 ,$$

$$f(r) = 0 + \tfrac{1}{2}\sigma^4 e^{-\sigma^2} \cdot r + o(r) ,$$

$$\frac{1}{r} f(r) \to \tfrac{1}{2}\sigma^4 e^{-\sigma^2} \quad \text{as} \quad r \to 0 ,$$

$$(n - k)[(1 - \frac{\sigma^2}{n - k})^{n-k} - (1 - \frac{\sigma^2}{(n - k)/2})^{(n-k)/2}] \to \tfrac{1}{2}\sigma^4 e^{-\sigma^2} \quad \text{as} \quad n \to +\infty,$$

$$(n - k) \text{ Var } (e^{\hat{\sigma}^2/2}) \to \frac{(\sigma^4/2)e^{-\sigma^2}}{e^{-2\sigma^2}} = \tfrac{1}{2}\sigma^4 e^{\sigma^2} \quad \text{as} \quad n \to +\infty .$$

Since k is fixed, we have

$$n \text{ Var } (e^{\hat{\sigma}^2/2}) \to \tfrac{1}{2}\sigma^4 e^{\sigma^2} \quad \text{as} \quad n \to +\infty . \qquad ||$$

(iii) $$E\ (\frac{1}{n} \sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) \to e^{\sigma^2/2} .$$

The least squares residual is distributed as

$$\hat{\varepsilon}_i = \varepsilon_i - x_i(X'X)^{-1}X'\varepsilon \sim N(0,\sigma^2(1 - P_{ii})) ,$$

where $P_{ii} = x_i(X'X)^{-1}x_i'$ and $P = X(X'X)^{-1}X'$ = projection matrix for the linear subspace spanned by columns of X. It follows that

$$E\, e^{\hat{\varepsilon}_i} = e^{\frac{1}{2}\sigma^2(1-P_{ii})},$$

$$E\,(\frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) = e^{\frac{1}{2}\sigma^2} \cdot \frac{1}{n}\sum_{i=1}^{n} e^{-\frac{1}{2}\sigma^2 P_{ii}}.$$

Note that

$$1 \geq e^{-\frac{1}{2}\sigma^2 P_{ii}} \geq 1 - \tfrac{1}{2}\sigma^2 P_{ii},$$

$$1 \geq \frac{1}{n}\sum_{i=1}^{n} e^{-\frac{1}{2}\sigma^2 P_{ii}} \geq 1 - \tfrac{1}{2}\sigma^2 \cdot \frac{1}{n}\sum_{i=1}^{n} P_{ii} = 1 - \tfrac{1}{2}\sigma^2 \cdot \frac{k}{n},$$

where

$$k = \text{rank } X = \text{tr } P = \sum_{i=1}^{n} P_{ii}.$$

Therefore,

$$e^{\sigma^2/2} \geq E\,(\frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) \geq (1 - \tfrac{1}{2}\sigma^2 \cdot \frac{k}{n}) \cdot e^{\sigma^2/2},$$

$$E\,(\frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) = e^{\sigma^2/2} + O(\tfrac{1}{n}) \to e^{\sigma^2/2} \qquad (k \text{ fixed}). \quad \|$$

(iv) $$n \text{ Var } (\frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) \to e^{\sigma^2}(e^{\sigma^2} - 1 - \sigma^2).$$

Note that

$$\hat{\varepsilon}_i + \hat{\varepsilon}_j = \varepsilon_i + \varepsilon_j - x_i(X'X)^{-1}\varepsilon - x_j(X'X)^{-1}\varepsilon$$

$$\sim \begin{cases} N(0,4\sigma^2(1 - P_{ii})) & (i = j) \\ N(0,2\sigma^2(1 - \frac{P_{ii} + P_{jj}}{2} - P_{ij})) & (i \neq j) , \end{cases}$$

where $P_{ij} = x_i(X'X)^{-1}x_j'$ and $P = X(X'X)^{-1}X'$. It follows that

$$\text{Var}\,(e^{\hat{\varepsilon}_i}) = e^{\sigma^2(1-P_{ii})}[e^{\sigma^2(1-P_{ii})} - 1] ,$$

$$\text{Cov}\,(e^{\hat{\varepsilon}_i}, e^{\hat{\varepsilon}_j}) = e^{\sigma^2[1-(P_{ii}+P_{jj})/2]}[e^{-\sigma^2 P_{ij}} - 1] \qquad (i \neq j) .$$

$$\text{Var}\,(\frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) = \frac{1}{n^2}\sum_{i=1}^{n} e^{\sigma^2(1-P_{ii})}[e^{\sigma^2(1-P_{ii})} - 1]$$

$$+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i} e^{\sigma^2[1-(P_{ii}+P_{jj})/2]}[e^{-\sigma^2 P_{ij}} - 1] . \qquad \text{(II.1)}$$

We will refer the first term in (II.1) as the diagonal term and the second term as the off-diagonal term.

Using the bounds

$$e^{\sigma^2} \geq e^{\sigma^2(1-P_{ii})} \geq e^{\sigma^2}(1 - \sigma^2 P_{ii})$$

and

$$e^{2\sigma^2} \geq e^{2\sigma^2(1-P_{ii})} \geq e^{2\sigma^2}(1 - 2\sigma^2 P_{ii}) ,$$

we can bound the diagonal term in (II.1) from above by

$$\frac{1}{n^2} \sum_{i=1}^{n} [e^{2\sigma^2} - e^{\sigma^2}(1 - \sigma^2 P_{ii})]$$

$$= \frac{1}{n} e^{\sigma^2}(e^{\sigma^2} - 1) + \sigma^2 e^{\sigma^2} \cdot \frac{1}{n^2} \sum_{i=1}^{n} P_{ii}$$

$$= \frac{1}{n} e^{\sigma^2}(e^{\sigma^2} - 1) + \sigma^2 e^{\sigma^2} \cdot \frac{k}{n^2}$$

and from below by

$$\frac{1}{n^2} \sum_{i=1}^{n} [e^{2\sigma^2}(1 - 2\sigma^2 P_{ii}) - e^{\sigma^2}]$$

$$= \frac{1}{n} e^{\sigma^2}(e^{\sigma^2} - 1) - 2\sigma^2 e^{-2\sigma^2} \cdot \frac{1}{n^2} \sum_{i=1}^{n} P_{ii}$$

$$= \frac{1}{n} e^{\sigma^2}(e^{\sigma^2} - 1) - 2\sigma^2 e^{2\sigma^2} \cdot \frac{k}{n^2} .$$

Therefore, the diagonal term in (II.1) can be expressed as follows:

$$\frac{1}{n^2} \sum_{i=1}^{n} e^{\sigma^2(1-P_{ii})} [e^{\sigma^2(1-P_{ii})} - 1]$$

$$= \frac{1}{n} e^{\sigma^2}(e^{\sigma^2} - 1) + O(\frac{1}{n^2}) .$$

For evaluating the off-diagonal term in (II.1), we will introduce the following notation:

$$r_{ij} = e^{-\sigma^2 P_{ij}} - 1 + \sigma^2 P_{ij} ,$$

$$s_{ij} = 1 - e^{-\sigma^2 [(P_{ii}+P_{jj})/2]} .$$

Note that

$$0 \leq s_{ij} \leq \sigma^2 \cdot \frac{P_{ii} + P_{jj}}{2} , \tag{II.2}$$

$$0 \leq r_{ij} = \frac{\sigma^4 P_{ij}^2}{2} + \frac{\sigma^6 P_{ij}^3}{6} + \cdots$$

$$\leq \frac{\sigma^4 P_{ij}^2}{2} + \frac{\sigma^6 |P_{ij}|^3}{6} + \cdots$$

$$= P_{ij}^2 \cdot \left\{ \frac{\sigma^4}{2} + \frac{\sigma^6 |P_{ij}|}{6} + \cdots \right\}$$

$$\leq P_{ij}^2 \cdot \left\{ \frac{\sigma^4}{2} + \frac{\sigma^6}{6} + \cdots \right\}$$

$$= P_{ij}^2 (e^{\sigma^2} - 1 - \sigma^2) . \tag{II.3}$$

In the next to the last line of (II.3), since P is a projection matrix, we have

$$PP' = P ,$$

$$\sum_{j=1}^{n} P_{ij}^2 = P_{ii} , \tag{II.4}$$

$$P_{ii}^2 \leq P_{ii} .$$

Thus

$$0 \leq P_{ii} \leq 1 \ . \tag{II.5}$$

Substituting (II.5) in (II.4), we have

$$P_{ij}^2 \leq P_{ii} \leq 1, \qquad |P_{ij}| \leq 1 \ .$$

The off-diagonal term in (II.1) can be re-expressed as follows:

$$\frac{1}{n^2} \sum_{i \neq j} e^{\sigma^2}(1 - s_{ij})(-\sigma^2 P_{ij} + r_{ij}) \ ,$$

which can be decomposed into four terms:

$$\begin{aligned} &- \sigma^2 e^{\sigma^2} \cdot \frac{1}{n^2} \sum_{i \neq j} P_{ij} + \sigma^2 e^{\sigma^2} \cdot \frac{1}{n^2} \sum_{i \neq j} P_{ij} s_{ij} \\ &+ e^{\sigma^2} \cdot \frac{1}{n^2} \sum_{i \neq j} r_{ij} - e^{\sigma^2} \cdot \frac{1}{n^2} \sum_{i \neq j} r_{ij} s_{ij} \ . \end{aligned} \tag{II.6}$$

The first term in (II.6) equals

$$-\sigma^2 e^{\sigma^2} \cdot \frac{1}{n^2} \left\{ \sum_{ij} P_{ij} - \sum_{i} P_{ii} \right\} = -\sigma^2 e^{\sigma^2} \cdot \frac{n - k}{n^2} \ ^*$$

$$= - \frac{1}{n} \sigma^2 e^{\sigma^2} + 0\left(\frac{1}{n^2}\right) \ .$$

---

*Assume X contains the intercept, then $\sum_{ij} P_{ij} = 1'P1 = 1'1 = n$.

By Cauchy-Schwarz inequality, the second term in (II.6) is dominated in absolute value by

$$\sigma^2 e^{\sigma^2} \cdot \frac{1}{n^2} \sqrt{\sum_{i\neq j} P_{ij}^2 \cdot \sum_{i\neq j} s_{ij}^2}$$

$$\leq \sigma^2 e^{\sigma^2} \cdot \frac{1}{n^2} \sqrt{\sum_{i\neq j} P_{ij}^2 \cdot \sum_{i\neq j} \left(\sigma^2 \cdot \frac{P_{ii} + P_{jj}}{2}\right)^2} \quad \text{[use (II.2)]}$$

$$\leq \frac{\sigma^4 e^{\sigma^2}}{2} \cdot \frac{1}{n^2} \sqrt{\sum_{i\neq j} P_{ij}^2 \cdot \sum_{i\neq j} (2P_{ii}^2 + 2P_{jj}^2)}$$

$$[(P_{ii} + P_{jj})^2 \leq 2P_{ii}^2 + 2P_{jj}^2]$$

$$= \frac{\sigma^4 e^{\sigma^2}}{2} \cdot \frac{1}{n^2} \sqrt{\sum_{i\neq j} P_{ij}^2 \cdot 4(n-1) \sum_{i} P_{ii}^2} \, .$$

Note that

$$\sum_{i\neq j} P_{ij}^2 \leq \sum_{ij} P_{ij}^2 = \sum_{i} \left(\sum_{j} P_{ij}^2\right) = \sum_{i} P_{ii} = k \, , \qquad \text{(II.7)}$$

$$\sum_{i} P_{ii}^2 \leq \sum_{i} P_{ii} = k \, . \qquad \text{(II.8)}$$

Therefore, the second term in (II.6) is dominated in absolute value by

$$\frac{\sigma^4 e^{\sigma^2}}{2} \cdot \frac{1}{n^2} \sqrt{4k^2(n-1)} = \frac{\sqrt{n-1}}{n^2} \cdot k\sigma^4 e^{\sigma^2} = O(\frac{1}{n^{3/2}}) \ .$$

The third term in (II.6) is dominated from below by zero and from above by

$$e^{\sigma^2} \cdot \frac{1}{n^2} \sum_{i \neq j} P_{ij}^2 (e^{\sigma^2} - 1 - \sigma^2) \leq \frac{k}{n^2} e^{\sigma^2} (e^{\sigma^2} - 1 - \sigma^2) = O(\frac{1}{n^2}) \ .$$

[use (II.7)]

Again using Cauchy-Schwarz inequality, the fourth term in (II.6) is dominated in absolute value by

$$e^{\sigma^2} \cdot \frac{1}{n^2} \sqrt{\sum_{i \neq j} s_{ij}^2 \cdot \sum_{i \neq j} r_{ij}^2}$$

$$\leq \sigma^2 e^{\sigma^2} (e^{\sigma^2} - 1 - \sigma^2) \frac{1}{n^2} \sqrt{\sum_{i \neq j} \left(\frac{P_{ii} + P_{jj}}{2} \right)^2 \cdot \sum_{i \neq j} P_{ij}^4}$$

[(II.2),(II.3)]

$$\leq \sigma^2 e^{\sigma^2} (e^{\sigma^2} - 1 - \sigma^2) \cdot \frac{1}{n^2} \sqrt{(n-1) \sum_{i} P_{ii}^2 \cdot \sum_{i \neq j} P_{ij}^2}$$

$[(P_{ii} + P_{jj})^2 \leq 2P_{ii}^2 + 2P_{jj}^2, \ P_{ij}^4 \leq P_{ij}^2]$

$$\leq \sigma^2 e^{\sigma^2} (e^{\sigma^2} - 1 - \sigma^2) \cdot \frac{1}{n^2} \sqrt{(n-1)k \cdot k} \qquad [(II.7),(II.8)]$$

$$= O(\frac{1}{n^{3/2}}) \ .$$

We have shown that the off-diagonal term in (II.1), i.e., the expression (II.6), equals

$$-\frac{1}{n}\sigma^2 e^{\sigma^2} + O(\frac{1}{n^{3/2}}) \ .$$

Combining the two terms in (II.1), we have

$$\text{Var}\ (\frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) = \frac{1}{n} e^{\sigma^2}(e^{\sigma^2} - 1 - \sigma^2) + O(\frac{1}{n^{3/2}})$$

$$n\ \text{Var}\ (\frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}) \to e^{\sigma^2}(e^{\sigma^2} - 1 - \sigma^2) \ . \quad ||$$

Addendum III
PROOF OF LEMMA 5

The convergence of the expectation follows from the convergence of the individual terms immediately.

The variance term can be decomposed as

$$n \operatorname{Var}\left(\frac{1}{J}\sum_{j=1}^{J} e^{\xi_j\hat{\beta}}\right) = \frac{1}{J^2}\sum_{j=1}^{J}\left[n \operatorname{Var}\left(e^{\xi_j\hat{\beta}}\right)\right] + \frac{1}{J^2}\sum_{j\neq k}\left[n \operatorname{Cov}\left(e^{\xi_j\hat{\beta}}, e^{\xi_k\hat{\beta}}\right)\right] . \tag{III.1}$$

The first term in (III.1) converges to

$$\frac{\sigma^2}{J^2}\sum_{j=1}^{J}(\xi_j\Psi^{-1}\xi_j')e^{2\xi_j\beta}$$

as individual terms converge to their limit given in Lemma 3. We are left with the covariance terms

$$n \operatorname{Cov}\left(e^{\xi_j\hat{\beta}}, e^{\xi_k\hat{\beta}}\right) = n\left[E\, e^{(\xi_j+\xi_k)\hat{\beta}} - E\, e^{\xi_j\hat{\beta}} \cdot E\, e^{\xi_k\hat{\beta}}\right]$$

$$= n\Big[e^{(\xi_j+\xi_k)\beta+(1/2n)\sigma^2(\xi_j+\xi_k)(X'X/n)^{-1}(\xi_j+\xi_k)'}$$

$$- e^{\xi_j\beta+(1/2n)\sigma^2\xi_j(X'X/n)^{-1}\xi_j'}$$

$$\cdot\, e^{\xi_k\beta+(1/2n)\sigma^2\xi_k(X'X/n)^{-1}\xi_k'}\Big]$$

$$= e^{(\xi_j+\xi_k)\beta+(1/2n)\sigma^2[\xi_j(X'X/n)^{-1}\xi_j'+\xi_k(X'X/n)^{-1}\xi_k']}$$

$$\cdot\, n\left[e^{(1/n)\sigma^2\xi_j(X'X/n)^{-1}\xi_k'} - 1\right]$$

$$\to e^{(\xi_j+\xi_k)\beta} \cdot \sigma^2\xi_j\Sigma^{-1}\xi_k' \,.$$

Therefore,

$$n \text{ Var } (\frac{1}{J}\sum_{j=1}^{J} e^{\xi_j\hat{\beta}}) \to \frac{\sigma^2}{J^2}\sum_{j=1}^{J}(\xi_j\Sigma^{-1}\xi_j')e^{2\xi_j\beta}$$

$$+ \frac{\sigma^2}{J^2}\sum_{j\neq k}(\xi_j\Sigma^{-1}\xi_k')e^{(\xi_j+\xi_k)\beta}$$

$$= \sigma^2 \cdot (\frac{1}{J}\sum_{j=1}^{J} e^{\xi_j\beta}\xi_j)\Sigma^{-1}(\frac{1}{J}\sum_{k=1}^{J} e^{\xi_k\beta}\xi_k)' \,.$$

Addendum IV
PROOF OF LEMMA 7

By the strong law of large numbers, as $J \to \infty$,

$$\text{(i)}\quad \frac{1}{J}\sum_j e^{\xi_j\beta} \to E\, e^{\xi\beta} \qquad \text{(a.s.)}$$

$$= \int \left(\frac{1}{\sqrt{2\pi}}\right)^k \cdot \frac{e^{\xi\beta}}{\det T} \cdot \exp\{-\tfrac{1}{2}(\xi - \bar{\xi})T^{-1}(\xi - \bar{\xi})'\}\, d\underset{\sim}{\xi}$$

$$= \int \left(\frac{1}{\sqrt{2\pi}}\right)^k \cdot \frac{1}{\det T} \cdot \exp\{-[\tfrac{1}{2}(\xi - \bar{\xi} - \beta'T)T^{-1}(\xi - \bar{\xi} - \beta'T)'$$

$$- \tfrac{1}{2}\beta'T\beta - \bar{\xi}\beta]\}\, d\underset{\sim}{\xi}$$

$$= \exp(\tfrac{1}{2}\beta'T\beta + \bar{\xi}\beta)\ .$$

$$\text{(ii)}\quad \frac{1}{J}\sum_j e^{\xi_j\beta}\xi_j \to E\, e^{\xi\beta}\xi \qquad \text{(a.s.)}$$

$$= \int \left(\frac{1}{\sqrt{2\pi}}\right)^k \frac{e^{\xi\beta}\xi}{\det T} \cdot \exp\{-\tfrac{1}{2}(\xi - \bar{\xi})T^{-1}(\xi - \bar{\xi})'\}\, d\underset{\sim}{\xi}$$

$$= \exp(\tfrac{1}{2}\beta'T\beta + \bar{\xi}\beta)$$

$$\cdot \int \left(\frac{1}{\sqrt{2\pi}}\right)^k \frac{1}{\det T}$$

$$\cdot\ \xi \exp\{-\tfrac{1}{2}(\xi - \overline{\xi} - \beta'T)T^{-1}(\xi - \overline{\xi}\beta'T)'\}\ d\underset{\sim}{\xi}$$

$$= (\overline{\xi} + \beta'T) \cdot \exp(\tfrac{1}{2}\beta'T\beta + \overline{\xi}\beta) \ .$$

Thus

$$\omega \to \left(\frac{E\ e^{\xi\beta}\xi}{E\ e^{\xi\beta}}\right)\ddagger^{-1}\left(\frac{E\ e^{\xi\beta}\xi}{E\ e^{\xi\beta}}\right)' = (\overline{\xi} + \beta'T)\ddagger^{-1}(\overline{\xi} + \beta'T)' \ . \qquad \text{(a.s.)}$$

REFERENCES TO APPENDIX B

Cox, D. R., and D. V. Hinkley, "A Note on the Efficiency of Least Squares Estimates," *Journal of Royal Statistical Society*, Series B, 30:284-289, 1968.

Duan, Naihua, "Consistency of Residual Distribution Function," 1980, unpublished manuscript.

Neyman, Jerzy, and Elizabeth Scott, "Correction for Bias Introduced by a Transformation of Variables," *Annals of Mathematical Statistics*, 31:643-655, 1960.

Appendix C

CORRECTING FOR INTRACLUSTER CORRELATION IN PROBIT REGRESSION MODELS

by *Naihua Duan*

C.1. INTRODUCTION

For a variety of problems, it is desirable to relate a dichotomous response (Yes or No) to some given explanatory variables, and to carry out certain inferences about the fitted relationships. A probit regression model is often used for this purpose.

An important restriction of the usual probit regression model is that the individual observations have to be stochastically independent. Although this is a reasonable assumption for many applications, it is unlikely to be true for a large class of studies of social data. For example, if households are sampled, but individuals are observed, then it is very likely that individual observations in the same cluster (household) will be correlated because of shared characteristics *not* observed as explanatory variables.

In the absence of stochastic independence among the individual observations, the estimated regression coefficients might still be statistically consistent. However, the estimated standard errors can be severely biased. Therefore it is necessary to consider modifications of the usual probit regression model to account for the possible lack of independence.

The usual probit regression model can be defined as follows:

$$P(Z = 1) = \Phi(x\beta) \ ,$$

$$P(Z = 0) = \overline{\Phi}(x\beta) \ ,$$

where $Z = 0, 1$ is the observed dichotomous response, Φ is the standard normal c.d.f., $\overline{\Phi}$ is its complement $1 - \Phi$, x is a given row vector of known characteristics, and β is the column vector of unknown regression parameters.

For convenience, we usually regard the probit regression model as a dichotomized continuous regression model:

$$
\begin{aligned}
&Y = -x\beta + \varepsilon \ , \\
&\varepsilon \sim N(0, 1) \quad \text{independent}, \\
&Z = 1 \qquad \text{if} \quad Y \leq 0 \ , \\
&\ \ = 0 \qquad \text{if} \quad Y > 0 \ ,
\end{aligned}
$$

where the *underlying propensity* Y is not observable. We will refer to this model as the *univariate probit regression model.*

The dependence among the individual observations in the same cluster can be described by modelling their joint distribution, or by modelling the joint distribution of their underlying propensities:

$$
\begin{aligned}
&Y_{ij} = -x_{ij}\beta + \varepsilon_{ij} \ , \quad j = 1, \ldots, J_i; \ i = 1, \ldots, I \ , \\
&\underset{\sim}{\varepsilon}_i \sim N(0, G_i) \ , \qquad \text{independent}, \\
&Z_{ij} = 1 \ , \qquad \text{if} \quad Y_{ij} \leq 0 \ , \\
&\ \ = 0 \ , \qquad \text{if} \quad Y_{ij} > 0 \ ,
\end{aligned}
$$

where G_i denotes the covariance matrix for the ith cluster. We will refer to this model as the *multivariate probit regression model.*

In this report, we will restrict our consideration to symmetric intraclass covariance matrices:[1]

---

[1]The symmetric intraclass assumption is plausible when the individuals in the same cluster are symmetric or exchangeable (e.g., if the cluster consists of students in the same class). It is less plausible for asymmetric clusters such as households in which the individuals do not play exchangeable roles--parents, children, etc. However, further complication in the covariance structure would make the problem less tractable, and so we will stay with the symmetric intraclass specification as an approximation even when the symmetry assumption might not be plausible.

$$G_i = \begin{bmatrix} 1 & & \rho \\ & \ddots & \\ \rho & & 1 \end{bmatrix} ,$$

$$(G_i)_{jj} \equiv 1 ,$$

$$(G_i)_{jk} \equiv \rho , \qquad j \neq k ,$$

where ρ denotes the constant correlation among ε_{ij}'s.[2]

In the discussion to follow, we will assume that the true value for the intracluster correlation ρ is known. We will discuss the estimation of ρ in Section C.6.

In principle, we should estimate the model and carry out inferences by using the full multivariate specification. However, the multivariate probit model requires the very expensive computation of multivariate normal integrals. Therefore, we choose to carry out the estimation of β with the computationally cheaper univariate specification.

By using the univariate specification to estimate the regression parameters, we will misstate the precision of the estimated coefficients--the estimated standard errors and t-statistics are based on the independence assumption, not accounting for the "loss of sample size" due to the intracluster correlation.

In Section C.2, we conjecture without proof that a generalization of Huber's result (Huber 1967)[3] can be applied to our problem, which provides that the incorrect univariate specification will lead to a consistent estimate of the regression parameters, together with a general formula for the precision of the estimated regression coefficients. In Section C.3, we derive the specific form of Huber's formula for the probit regression problem, and in Section C.4, we derive

---

[2]Because the scale of the underlying propensity is not identified, it can be standardized without loss of generality.

[3]References are cited at the end of the appendix.

a "fudge factor," or approximation, which can be used to adjust the precision estimated from the univariate probit model. In Section C.5, we discuss the application of this fudge factor to randomly assigned experimental treatments. In Section C.6, we discuss the problem of estimating the intracluster correlation ρ, and in Section C.7, we summarize our findings.

C.2. CONJECTURED GENERALIZATION OF HUBER'S RESULT

Huber (1967) examined the problem of maximizing a misspecified likelihood criterion $L_x(\theta)$ and proved that under certain regularity conditions the resulting misspecified maximum likelihood estimate $\hat{\theta}$ is consistent and asymptotically normal, with asymptotic covariance matrix

$$\Lambda^{-1}\Gamma\Lambda^{-1} ,$$

where Λ denotes the Fisher information matrix under the misspecified model and Γ denotes the true covariance matrix of the misspecified score function

$$S(\theta) = \frac{\partial}{\partial\theta} \log L(\theta) .$$

Huber (1967) proved this result for the one population case, where the observations are independently and identically distributed. Similar results have been proven for other cases: Huber (1973) and Bickel (1975) for linear regression, and Ogata (1980) for incorrect Markov models.

For the multivariate probit model, the clustered observations

$$\underset{\sim}{Z}_i = (Z_{i1}, \ldots, Z_{iJ_i})'$$

are stochastically independent; however, they are not identically distributed: the cluster sizes J_i and the observed characteristics $\{x_{ij}: j = 1, \ldots, J_i\}$ can be different for different clusters. We

conjecture that the generalization of Huber's result is still true for our problem. I am working to provide a rigorous proof of the conjecture.

C.3. ASYMPTOTIC COVARIANCE MATRICES

Under the univariate probit model, we have the misspecified likelihood function

$$L(\beta) = \prod_{i=1}^{I} L_i(\beta) = \prod_{i=1}^{I} \prod_{j=1}^{J_i} \Phi[(-1)^{1-Z_{ij}} x_{ij}\beta]$$

and the misspecified log-likelihood function

$$\ell(\beta) = \sum_{i=1}^{I} \ell_i(\beta) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} \log \Phi[(-1)^{1-Z_{ij}} x_{ij}\beta] \ .$$

The misspecified score function is

$$\nabla\ell(\beta) = \sum_{i=1}^{I} \nabla\ell_i(\beta) = \sum_{i=1}^{I} \sum_{j=1}^{J_i} x_{ij}' C_{ij}(\beta) \ ,$$

where

$$C_{ij}(\beta) = (-1)^{1-Z_{ij}} \phi(x_{ij}\beta) / \Phi[(-1)^{1-Z_{ij}} x_{ij}\beta]$$

ϕ = standard normal p.d.f.

The misspecified Fisher information matrix, evaluated at the true parameters value $\beta = \beta_o$, is

$$\Lambda(\beta_o) = E\,[-\nabla\nabla'\ell(\beta_o)]$$

$$= E\,[-\sum_{ij} \nabla C_{ij}(\beta_o) \cdot x_{ij}]$$

$$= E \sum_{ij} \left[(-1)^{1-Z_{ij}x_{ij}\beta_o} \cdot \frac{\phi(x_{ij}\beta_o)}{\Phi[(-1)^{1-Z_{ij}x_{ij}\beta_o}]} + \frac{\phi^2(x_{ij}\beta_o)}{\Phi^2[(-1)^{1-Z_{ij}x_{ij}\beta_o}]} \right]$$

$$\cdot\, x'_{ij}x_{ij}$$

$$= 0 + \sum_{i=1}^{I} \sum_{j=1}^{J_i} \frac{\phi^2(x_{ij}\beta_o)}{\Phi(x_{ij}\beta_o) \cdot \bar{\Phi}(x_{ij}\beta_o)} x'_{ij}x_{ij} \ ,$$

where $\bar{\Phi}$ denotes $1 - \Phi$. Note that the above expectation is the same for both the univariate and the multivariate probit models.

To simplify notation, let

$$\phi_{ij} = \phi(x_{ij}\beta_o) \ ,$$

$$\Phi_{ij} = \Phi(x_{ij}\beta_o) \ ,$$

$$\sigma^2_{ij} = \frac{\phi^2_{ij}}{\Phi_{ij} \cdot \bar{\Phi}_{ij}} \ ,$$

$$X = [x'_{11}, \ldots, x'_{1J_1}, \ldots, x'_{ij}, \ldots, x'_{IJ_I}]' \ ,$$

$$\Sigma = \text{diag}\,[\sigma_{11}, \ldots, \sigma_{1J_1}, \ldots, \sigma_{ij}, \ldots, \sigma_{IJ_I}] \ ,$$

where diag[...] denotes the diagonal matrix with the specified diagonal elements.

We have

$$\Lambda(\beta_o) = \sum_{ij} \sigma_{ij}^2 x_{ij}' x_{ij} = X' \ddagger^2 X \ . \qquad \text{(C.3.1)}$$

Without accounting for the intracluster correlation, we would have used

$$V_u(\beta_o) = \Lambda(\beta_o)^{-1} = (X' \ddagger^2 X)^{-1} = [(X' \ddagger^2 X)^{-1} X' \ddagger'] I [\ddagger X (X' \ddagger^2 X)^{-1}] \ , \qquad \text{(C.3.2)}$$

or an estimated version of it, e.g., $V_u(\hat{\beta})$, as the approximate covariance matrix for the estimated regression coefficients $\hat{\beta}$. As we have noted in the previous section, this will likely misstate the precision of $\hat{\beta}$.

We will compute the asymptotically correct covariance matrix according to Huber's formula.

The covariance matrix of the misspecified score function, evaluated for the true parameter value $\beta = \beta_o$, is

$$\begin{aligned} \Gamma(\beta_o) &= E\,[\nabla \ell(\beta_o)][\nabla \ell(\beta_o)]' \\ &= \sum_{i=1}^{I} E\,[\nabla \ell_i(\beta_o)][\nabla \ell_i(\beta_o)]' \\ &= \sum_{i=1}^{I} E\,[\sum_{j=1}^{J_i} x_{ij}' C_{ij}(\beta_o)][\sum_{j=1}^{J_i} x_{ij}' C_{ij}(\beta_o)]' \\ &= \sum_{i=1}^{I} \sum_{jk} x_{ij}' d_{ij,ik} x_{ik} \ , \end{aligned}$$

where

$$d_{ij,ij} = E\, C_{ij}^2(\beta_o) = \sigma_{ij}^2 ,$$

$$d_{ij,ik} = E\, C_{ij}(\beta_o) C_{ik}(\beta_o)$$

$$= \phi_{ij}\phi_{ik}\left[\frac{\Phi 2(x_{ij}\beta_o, x_{ik}\beta_o)}{\Phi_{ij} \cdot \Phi_{ik}} - \frac{\Phi 2(x_{ij}\beta_o, -x_{ik}\beta_o)}{\Phi_{ij} \cdot \overline{\Phi}_{ik}}\right.$$

$$\left. - \frac{\Phi 2(-x_{ij}\beta_o, x_{ik}\beta_o)}{\overline{\Phi}_{ij} \cdot \Phi_{ik}} + \frac{\Phi 2(-x_{ij}\beta_o, -x_{ik}\beta_o)}{\overline{\Phi}_{ij} \cdot \overline{\Phi}_{ik}}\right]$$

$$= \phi_{ij} \cdot \phi_{ik} \cdot \frac{\Phi 2_{ij,ik} - \Phi_{ij} \cdot \Phi_{ik}}{\Phi_{ij} \cdot \overline{\Phi}_{ij} \cdot \Phi_{ik} \cdot \overline{\Phi}_{ik}} ,$$

$$= r_{ij,ik}\sigma_{ij}\sigma_{ik} , \qquad [j \neq k] ,$$

where $\Phi 2$ is the standard bivariate normal c.d.f. with correlation ρ, $\Phi 2_{ij,ik}$ is $\Phi 2(x_{ij}\beta_o, x_{ik}\beta_o)$, and

$$r_{ij,ik} = \text{Corr}\,(z_{ij}, z_{ik}) = \frac{\Phi 2_{ij,ik} - \Phi_{ij} \cdot \Phi_{ik}}{\sqrt{\Phi_{ij} \cdot \overline{\Phi}_{ij} \cdot \Phi_{ik} \cdot \overline{\Phi}_{ik}}} .$$

We will further introduce the notation

$$R = \text{DIAG}\,[R_1, \ldots, R_i, \ldots, R_I] ,$$

$$(R_i)_{jj} = 1 ,$$

$$(R_i)_{jk} = r_{ij,ik} , \qquad (j \neq k)$$

where DIAG [...] creates the block diagonal matrix with the specified diagonal blocks.

We have

$$\Gamma(\beta_o) = \sum_{ij} \sigma_{ij}^2 x_{ij}' x_{ij} + \sum_i \sum_{j \neq k} r_{ij,ik} \sigma_{ij} \sigma_{ik} x_{ij}' x_{ik} = X' \ddagger' R \ddagger X \ .$$

The asymptotic covariance matrix can be expressed as

$$V_m(\beta_o) = \Lambda(\beta_o)^{-1} \Gamma(\beta_o) \Lambda(\beta_o)^{-1} = [(X' \ddagger^2 X)^{-1} X' \ddagger'] R [\ddagger X (X' \ddagger^2 X)^{-1}] \ . \tag{C.3.3}$$

Note that the only difference between $V_u(\beta_o)$ and $V_m(\beta_o)$ [Eqs. (C.3.2) and (C.3.3)] is the replacement of the covariance matrix R by the identity matrix. Obviously, if there is no intracluster correlation, the two formulas will be identical.

C.4. FUDGE FACTOR

The asymptotically correct covariance matrix $V_m(\beta_o)$ can be estimated with specially designed software after estimating the univariate probit model. However, for many purposes, it will be sufficient to estimate (or bound) the effect of intracluster correlation on the precision (standard error or t-statistic) estimated from the univariate probit model, and make the appropriate adjustment. For this purpose we will develop a fudge factor to make an approximate adjustment.

We make the following conjecture:

Conjecture C.4.1. If the intracluster correlation ρ is nonnegative, we can bound $r_{ij,ik} = \text{Corr}\ (Z_{ij}, Z_{ik})$ from above by

$$r^* = 4 \cdot \Phi 2(0,0) - 1 \ . \tag{C.4.1}$$

Note that

$$r^* = \frac{\Phi 2(0,0) - \Phi(0) \cdot \Phi(0)}{\sqrt{\Phi(0) \cdot \bar{\Phi}(0) \cdot \Phi(0) \cdot \bar{\Phi}(0)}}$$

$$= \text{Corr}\ (Z_{ij}, Z_{ik};\ x_{ij}\beta_o = x_{ik}\beta_o = 0)\ . \qquad \|$$

We don't have a complete mathematical proof of the conjecture. A partial proof is given below, which shows that "$x_{ij}\beta_o = x_{ik}\beta_o = 0$" is a critical point of $r_{ij,ik}(x_{ij}\beta_o, x_{ik}\beta_o)$. What is missing is that the critical point might not be unique, and might not be the global maximum.

Partial Proof. Let t,s denote the thresholds. The correlation of interest has the expression

$$r = \frac{\Phi 2(t,s) - \Phi(t) \cdot \Phi(s)}{\sqrt{\Phi(t) \cdot \bar{\Phi}(t) \cdot \Phi(s) \cdot \bar{\Phi}(s)}}\ .$$

(1) The denominator assumes its maximum value when $t = s = 0$.

(2) The numerator has gradient vector

$$\frac{\partial(\text{num})}{\partial r} = \phi(t) \cdot \Phi\left(\frac{s - \rho t}{\sqrt{1 - \rho^2}}\right) - \phi(t) \cdot \Phi(s)\ ,$$

$$\frac{\partial(\text{num})}{\partial s} = \phi(s) \cdot \Phi\left(\frac{t - \rho s}{\sqrt{1 - \rho^2}}\right) - \phi(s) \cdot \Phi(t)\ .$$

The only solution to the gradient equations

$$\frac{\partial(\text{num})}{\partial t} = 0\ , \qquad \frac{\partial(\text{num})}{\partial s} = 0$$

is $t = s = 0$.

(3) The Hessian matrix of the numerator at $t = s = 0$ is

$$-2\pi \begin{bmatrix} \dfrac{\rho}{\sqrt{1-\rho^2}} & 1 - \dfrac{1}{\sqrt{1-\rho^2}} \\ 1 - \dfrac{1}{\sqrt{1-\rho^2}} & \dfrac{\rho}{\sqrt{1-\rho^2}} \end{bmatrix} ,$$

which is negative definite for $\rho \geq 0$. Therefore, the numerator assumes a local maximum at $t = s = 0$.

(4) The gradient vector for $r(t,s)$ will vanish if both gradient vectors for the denominator and numerator vanish, e.g., if $t = s = 0$. Therefore, r has a critical point at $t = s = 0$. ||

In order to supplement the incomplete mathematical proof, we have carried out a numerical evaluation, with results given as Tables C.4.1, C.4.2, and C.4.3. In each of the tables, we have evaluated $r_{ij,ik}(x_{ij}\beta_o, x_{ik}\beta_o)$ for various thresholds $x_{ij}\beta_o$ and $x_{ik}\beta_o$ (denoted as S_j and S_k in the tables).

For the three levels of intracluster correlation considered ($\rho = .35, .45, .55$), $r^* = r(0,0)$ is indeed the maximum among all $r(S_j,S_k)$'s tabulated. (Here, $r^* = r(0,0)$ corresponds to the underlined entry "$\Phi(S_j) = .50$, $\Phi(S_k) = .50$" in the tables.) Moreover, even if the conjecture is false, r^* still serves as a reasonable approximation for $r(S_j,S_k)$ over a wide range of (S_j,S_k)'s.

Based on the conjecture, we have

Lemma C.4.2. If the intracluster correlation ρ is nonnegative, we can bound R from above by

$$R^* = \text{DIAG}\ [R_1^*, \ldots, R_i^*, \ldots, R_I^*] ,$$

$$(R_i^*)_{jj} = 1 ,$$

$$(R_i^*)_{jk} = r^* , \qquad (j \neq k)$$

in the sense that the difference $R^* - R$ is nonnegative (*not* necessarily nonnegative definite, though). ||

Table C.4.1

CORRELATION OF DISCRETE DECISIONS,* $\rho = .35$

$\Phi(S_k)$ \ r** \ $\Phi(S_j)$	.10	.20	.30	.40	.50	.60	.70	.80	.90
.10	.156								
.20	.171	.194							
.30	.173	.202	.215						
.40	.169	.202	.218	.225					
.50	.161	.197	.215	.225	.228				
.60	.149	.186	.207	.219	.224	.224			
.70	.135	.171	.194	.207	.215	.217	.213		
.80	.115	.150	.172	.187	.196	.201	.201	.194	
.90	.086	.115	.136	.151	.161	.168	.172	.171	.157

*We have omitted the upper triangular part of the table because the correlation r is symmetric in its two arguments.

**
$$r = r(S_j, S_k) = \frac{\Phi 2(S_j, S_k) - \Phi(S_j) \cdot \Phi(S_k)}{\sqrt{\Phi(S_j) \cdot \bar{\Phi}(S_j) \cdot \Phi(S_k) \cdot \bar{\Phi}(S_k)}} .$$

Table C.4.2

CORRELATION OF DISCRETE DECISIONS,* $\rho = .45$

$\Phi(S_k)$ \ r** \ $\Phi(S_j)$	.10	.20	.30	.40	.50	.60	.70	.80	.90
.10	.216								
.20	.230	.260							
.30	.228	.267	.283						
.40	.219	.264	.286	.294					
.50	.204	.254	.280	.293	.298				
.60	.186	.237	.267	.284	.293	.293			
.70	.164	.214	.246	.267	.279	.284	.281		
.80	.136	.183	.215	.238	.253	.263	.266	.259	
.90	.098	.137	.165	.187	.205	.218	.227	.230	.216

*We have omitted the upper triangular part of the table because the correlation r is symmetric in its two arguments.

**
$$r = r(S_j, S_k) = \frac{\phi 2(S_j, S_k) - \phi(S_j) \cdot \phi(S_k)}{\sqrt{\phi(S_j) \cdot \overline{\phi}(S_j) \cdot \phi(S_k) \cdot \overline{\phi}(S_k)}} .$$

Table C.4.3

CORRELATION OF DISCRETE DECISIONS,* $\rho = .55$

$\Phi(S_k)$ \ r** \ $\Phi(S_j)$	.10	.20	.30	.40	.50	.60	.70	.80	.90
.10	.284								
.20	.296	.332							
.30	.287	.337	.356						
.40	.268	.329	.357	.368					
.50	.245	.311	.347	.365	.371				
.60	.218	.286	.327	.352	.365	.366			
.70	.188	.253	.296	.327	.346	.356	.354		
.80	.152	.211	.253	.286	.311	.328	.336	.331	
.90	.106	.152	.189	.220	.245	.267	.286	.296	.285

*We have omitted the upper triangular part of the table because the correlation r is symmetric in its two arguments.

**
$$r = r(S_j, S_k) = \frac{\phi 2(S_j, S_k) - \phi(S_j) \cdot \phi(S_k)}{\sqrt{\overline{\phi}(S_j) \cdot \overline{\phi}(S_j) \cdot \overline{\phi}(S_k) \cdot \overline{\phi}(S_k)}} .$$

We can express R^* by

$$R^* = (1 - r^*)I + r^*D ,$$

where

$$D = \text{DIAG } [D_1, \ldots, D_i, \ldots, D_I] ,$$

$$(D_i)_{jk} \equiv 1 .$$

Replacing R by R^*, we have the expression for covariance matrix

$$V_m^*(\beta_o) = (1 - r^*)V_u(\beta_o) + r^*[(X'\ddagger^2X)^{-1}X'\ddagger']D[\ddagger X(X'\ddagger^2X)^{-1}] .$$

We can either regard $V_m^*(\beta_o)$ as an approximation for $V_m(\beta_o)$, or as an upper bound in the sense of Lemma C.4.2.

Denoting the kth row of $(X'\ddagger^2X)^{-1}X'\ddagger'$ by $b^{(k)}$, we have the expression for the asymptotic variance of $\hat{\beta}_k$ as

$$[V_m^*(\beta_o)]_{kk} = b^{(k)}R^*b^{(k)'}$$

$$= (1 - r^*)b^{(k)}b^{(k)'} + r^*b^{(k)}Db^{(k)'}$$

$$= (1 - r^*) \sum_i \sum_j (b_{ij}^{(k)})^2 + r^* \sum_i (\sum_j b_{ij}^{(k)})^2 .$$

Again, we can regard $[V_m^*(\beta_o)]_{kk}$ as an approximation for $[V_m(\beta_o)]_{kk}$. If the coefficients $b_{ij}^{(k)}$ have the same signs in the same cluster, it is an upper bound.

The univariate probit model would use

$$[V_u(\beta_o)]_{kk} = b^{(k)}b^{(k)'} = \sum_i \sum_j (b_{ij}^{(k)})^2$$

as the variance of $\hat{\beta}_k$. The adjustment, therefore, is the multiplicative factor

$$\frac{[V_m(\beta_o)]_{kk}}{[V_u(\beta_o)]_{kk}} \approx \frac{[V_m^*(\beta_o)]_{kk}}{[V_u(\beta_o)]_{kk}}$$

$$= 1 - r^* + r^* \frac{\sum_i (\sum_j b_{ij}^{(k)})^2}{\sum_i \sum_j (b_{ij}^{(k)})^2}$$

$$= 1 - r^* + r^* \frac{\sum_i J_i^2 (\bar{b}_{i.}^{(k)})^2}{\sum_i J_i (\bar{b}_{i.}^{(k)})^2 + \sum_i J_i S_i^2}$$

$$\leq 1 - r^* + r^* \frac{\sum_i J_i^2 (\bar{b}_{i.}^{(k)})^2}{\sum_i J_i (\bar{b}_{i.}^{(k)})^2}, \qquad \text{(C.4.2)}$$

where

$$\bar{b}_{i.}^{(k)} = \frac{1}{J_i} \sum_j b_{ij}^{(k)},$$

$$S_i^2 = \frac{1}{J_i} \sum_i (b_{ij}^{(k)} - \bar{b}_{i.}^{(k)})^2 .$$

(The approximation $[V_m^*(\beta_o)]$, above, can be replaced by upper bound if the coefficients $b_{ij}^{(k)}$ have the same sign in the same cluster.)

If the cluster size J_i is approximately independent of $\bar{b}_{i.}^{(k)}$, we have

$$\frac{\sum_i J_i^2 (\bar{b}_i^{(k)})^2}{\sum_i J_i (\bar{b}_{i\cdot}^{(k)})^2} \approx \frac{\sum_i J_i^2}{\sum_i J_i} . \qquad (C.4.3)$$

Theorem C.4.3. If the intracluster correlation ρ is nonnegative, and if the cluster size J_i is approximately independent of $\bar{b}_{i\cdot}^{(k)}$, the estimated precision based on the univariate probit model can be adjusted multiplicatively by the fudge factor

$$1 - r^* + r^* \cdot \frac{\sum_i J_i^2}{\sum_i J_i} , \qquad (C.4.4)$$

where $r^* = 4 \cdot \Phi 2(0,0) - 1$ and $\Phi 2$ is the standard bivariate normal c.d.f. with correlation ρ. The fudge factor is an approximate upper bound in general, and an exact upper bound if the coefficients $b_{ij}^{(k)}$ have the same signs in the same cluster. If the cluster size J_i is appreciably dependent on $\bar{b}_i^{(k)}$, no such simple fudge factor exists; the approximate upper bound (C.4.3) has to be used. ||

For the Health Insurance Study population, we have

$$\frac{\sum_i J_i^2}{\sum_i J_i} = 3.87 .$$

For the dichotomous decision to use medical care (first equation in the two-part and four-part models), we estimated the intracluster (intrafamily) correlation ρ to be .45, using a method described in the last section. The corresponding r^* from Table C.4.2 is .30. Therefore, the adjustment in Theorem C.4.3 is $1.86 = (1.36)^2$. The standard errors estimated from the univariate probit regression model should be inflated by at most 36 percent.

C.5. EXPERIMENTAL TREATMENT

For many purposes, the most crucial inferences are comparisons of experimental treatments. If the treatments are assigned *randomly* to the individual clusters, with individuals in the same clusters assigned to the same treatment, the column vectors $X^{(k)}$ for the treatments will be approximately orthogonal to the other explanatory variables, and the coefficient $X_{ij}^{(k)}$ will be the same (thus have the same signs) in the same cluster. Moreover, the cluster size J_i will be approximately independent of $\bar{x}_{i.}^{(k)}$. Therefore the two requirements in Theorem C.4.3 for the fudge factor (C.4.3) to be an exact upper bound will be satisfied for those variables in terms of $X^{(k)}$.

The vectors $b^{(k)}$ = kth row of $(X'\sharp^2X)^{-1}X'\sharp'$ are related to $X^{(k)}$ through (1) reweighting by $\sharp$ and (2) rotation by $(X'\sharp^2X)^{-1}$. (1) The reweighting is moderate. As $x_{ij}\beta_o$ ranges from 0 through 2, β_{ij} changes by about a factor of two (Table C.4.4). Such moderate reweighting is unlikely to change the qualitative features of $X^{(k)}$ drastically. (2) Because of the approximate orthogonality between the treatment variables and the other explanatory variables, the unweighted rotation matrix $(X'X)^{-1}$ will be approximately block diagonal, with one block corresponding to the treatment variables (including the intercept) and another block corresponding to the other explanatory variables. Since the reweighting is moderate, it is unlikely to change the approximate orthogonality. Thus, the weighted rotation matrix $(X'\sharp^2X)^{-1}$ will still be approximately block diagonal. Therefore, the rotation for the treatment variables $X^{(k)}$ is almost entirely among themselves (including the intercept). Since all the treatment variables $X^{(k)}$ satisfy the conditions for Eq. (C.4.3) being an exact upper bound, rotation among the variables is not going to change these features.

Corollary C.5.1. For randomly assigned experimental treatments, if individuals in the same cluster are assigned to the same treatment, the fudge factor in (C.4.4) will approximately satisfy the conditions in Theorem C.4.3 for being an exact upper bound. ||

Table C.4.4

THE MAGNITUDE OF σ_{ij}

$x_{ij}\beta_o$	$\sigma_{ij} = \phi_{ij} / \sqrt{\Phi_{ij} \cdot \bar{\Phi}_{ij}}$
0.0	.798
± 0.5	.762
± 1.0	.662
± 1.5	.520
± 2.0	.360

C.6. ESTIMATION OF THE INTRACLUSTER CORRELATION

For either estimating the covariance matrix V_m in (C.3.3), or computing the adjustment (C.4.2) or the fudge factor (C.4.4), we need to estimate the intracluster correlation ρ for the underlying propensities.

To avoid the prohibitive computation of multivariate normal c.d.f., we can use a random subsample of two individuals from each cluster with two or more individuals. (The computation of bivariate normal c.d.f. is still expensive, but reasonably affordable.) Therefore, we have the reduced sample

$$\{(z_{ik}, x_{ik}) : k = 1,2;\ i = 1, \ldots, I^*\},$$

where $I^* \leq I$. (Single-member clusters are dropped by this procedure.) The likelihood function is

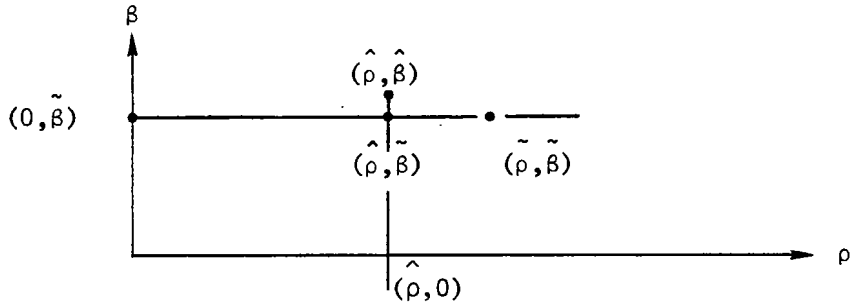
$$L(\rho,\beta) = \prod_{i=1}^{I^*} L_i(\rho,\beta)$$

$$= \prod_{i=1}^{I^*} \Phi 2[(-1)^{1-z_{i1}} x_{i1}\beta, (-1)^{1-z_{i2}} x_{i2}\beta;\ \rho],$$

where $\Phi 2(*,*;\rho)$ denotes the standard bivariate normal c.d.f. with correlation ρ.

Lemma C.6.1. Let $\tilde{\beta}$ be a consistent estimate of β (e.g., the univariate probit regression estimate). We can obtain a consistent estimate of ρ by maximizing $L(\rho,\tilde{\beta})$. ‖

Proof. We will denote the maximum likelihood estimate of (ρ,β) by $(\hat{\rho},\hat{\beta})$, the given consistent estimate of β by $\tilde{\beta}$, the estimate maximizing $L(\rho,\tilde{\beta})$ by $\tilde{\rho}$. We will also denote the log-likelihood by ℓ, the partial differentiation in ρ by ∇_ρ, and the partial differentiation in β by ∇_β:



Expand the function $\nabla_\rho \ell(\rho,\tilde{\beta})$ in ρ near $\rho = \hat{\rho}$:

$$\nabla_\rho \ell(\rho,\tilde{\beta}) = \nabla_\rho \ell(\hat{\rho},\tilde{\beta}) + (\rho - \hat{\rho}) \cdot \nabla_\rho^2 \ell(\hat{\rho},\tilde{\beta}) + \ldots .$$

Since $\tilde{\rho}$ maximizes $\ell(\rho,\tilde{\beta})$, the partial derivative $\nabla_\rho \ell(\rho,\tilde{\beta})$ vanishes at $\rho = \tilde{\rho}$.

$$0 = \nabla_\rho \ell(\hat{\rho},\tilde{\beta}) + (\tilde{\rho} - \hat{\rho}) \cdot \nabla_\rho^2 \ell(\hat{\rho},\tilde{\beta}) + \ldots$$

$$\tilde{\rho} - \hat{\rho} \approx \frac{\nabla_\rho \ell(\hat{\rho},\tilde{\beta})}{-\nabla_\rho^2 \ell(\hat{\rho},\tilde{\beta})} .$$

Expanding the function $\nabla_\rho \ell(\hat{\rho},\beta)$ in β near $\beta = \hat{\beta}$ yields

$$\nabla_\rho \ell(\hat{\rho},\beta) = \nabla_\rho \ell(\hat{\rho},\hat{\beta}) + (\beta - \hat{\beta})'\nabla_\beta \nabla_\rho \ell(\hat{\rho},\hat{\beta}) + \ldots .$$

Since $(\hat{\rho},\hat{\beta})$ maximizes $\ell(\rho,\beta)$, the partial derivative $\nabla_\rho \ell(\hat{\rho},\hat{\beta})$ vanishes. Therefore

$$\nabla_\rho \ell(\hat{\rho},\tilde{\beta}) = 0 + (\tilde{\beta} - \hat{\beta})'\nabla_\beta \nabla_\rho \ell(\hat{\rho},\hat{\beta}) .$$

Therefore

$$\tilde{\rho} - \hat{\rho} \approx \frac{(\tilde{\beta} - \hat{\beta})'\nabla_\beta \nabla_\rho \ell(\hat{\rho},\hat{\beta})}{-\nabla_\rho^2 \ell(\hat{\rho},\tilde{\beta})}$$

$$\approx - \frac{(\tilde{\beta} - \hat{\beta})' I(\rho,\beta)}{I(\rho,\rho)} ,$$

where $I(\rho,\beta)$ is the off-diagonal block in the average Fisher information matrix, and $I(\rho,\rho)$ is the diagonal term for ρ.

We have assumed $\tilde{\beta}$ to be consistent; under mild regularity conditions on the design matrix, $\hat{\beta}$ will also be consistent. Therefore $\tilde{\beta} - \hat{\beta}$ will converge to zero, and $\tilde{\rho} - \hat{\rho}$ will converge to zero as long as $I(\rho,\rho)$ is bounded away from zero. Again $\hat{\rho}$ will be consistent under mild regularity conditions, which proves that $\tilde{\rho}$ will be consistent. ||

The maximization required in Lemma C.6.1 can be carried out by a grid search. The numerical evaluation of the likelihood function would require the computation of bivariate normal c.d.f., which can be carried out either with numerical double integration or some series expansion methods.

C.7. CONCLUSION

Assuming the generalization of Huber's result conjectured in Section C.2, we derived the asymptotic covariance matrix (C.3.3) for the multivariate probit regression model. We also derived a fudge factor (C.4.4) that can be used to adjust the standard errors estimated from the univariate probit regression model. Under certain conditions, such as an analysis of experimental treatment, the fudge factor can be regarded as an upper bound.

REFERENCES TO APPENDIX C

Bickel, P. J., "One Step Huber Estimates in the Linear Model," *Journal of American Statistical Association*, Vol. 70, 1975, pp. 428-434.

Huber, P. J., "The Behavior of Maximum Likelihood Estimates Under Non-Standard Conditions," *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, I, 1967, pp. 221-233.

----, "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *Annals of Statistics*, Vol. 1, 1973, pp. 799-821.

Ogata, Yoshihiko, "Maximum Likelihood Estimates of Incorrect Mankov Models for Time Series and the Derivation of AIC," *Journal of Applied Probability*, Vol. 17, 1980, pp. 59-72.

Appendix D

PREDICTIVE RESULTS FROM OTHER SITE-YEARS FOR ALTERNATIVE MODELS

Table D.1

DAYTON YEAR 3: PREDICTIONS (STANDARD ERROR)

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	408.21 (46.77)	404.96 (74.43)	634.84 (89.24)	496.37 (55.67)	514.22 (41.57)
P25	387.99 (82.45)	365.85 (82.45)	379.53 (57.74)	313.84 (39.20)	371.34 (35.75)
P50	217.47 (39.76)	215.28 (97.58)	247.49 (42.85)	244.87 (36.52)	325.90 (38.67)
PFD	309.51 (57.16)	325.91 (78.58)	261.84 (38.99)	207.00 (26.24)	320.82 (33.12)
IDP	572.39 (336.64)	596.85 (133.22)	285.43 (64.10)	271.04 (53.88)	346.78 (43.02)

Table D.2

SEATTLE YEAR 1: PREDICTIONS (STANDARD ERROR)

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	361.94 (40.83)	357.26 (35.60)	492.40 (54.43)	379.00 (33.27)	382.08 (27.95)
P25	323.02 (60.23)	339.56 (46.68)	401.68 (54.96)	314.78 (35.28)	325.46 (28.31)
P50	--	--	--	--	--
PFD	244.31 (41.47)	249.74 (48.80)	258.24 (37.02)	217.09 (26.49)	278.26 (27.13)
IDP	245.30 (25.33)	233.33 (44.05)	306.47 (39.56)	263.27 (28.44)	328.76 (28.69)

Table D.3

SEATTLE YEAR 2: PREDICTIONS (STANDARD ERROR)

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	412.74 (91.56)	400.03 (78.52)	517.13 (60.25)	363.85 (32.27)	442.30 (35.20)
P25	529.12 (148.74)	541.58 (105.17)	385.41 (56.49)	311.93 (36.71)	359.12 (35.01)
P50	--	--	--	--	--
PFD	281.93 (57.07)	307.39 (112.73)	267.56 (41.75)	260.92 (34.85)	317.46 (33.21)
IDP	274.90 (36.02)	264.46 (98.85)	306.91 (42.46)	260.96 (29.70)	357.18 (34.80)

Table D.4

FITCHBURG YEAR 1: PREDICTIONS (STANDARD ERROR)

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	494.76	477.64	441.13	345.57	376.30
	(172.13)	(114.89)	(63.10)	(41.67)	(32.64)
P25	332.41	314.89	358.22	335.81	337.77
	(80.59)	(161.18)	(67.53)	(55.66)	(35.87)
P50	212.98	238.86	261.85	215.24	267.86
	(56.41)	(237.84)	(70.37)	(50.10)	(38.40)
PFD	295.13	292.37	214.44	181.44	285.07
	(86.02)	(175.00)	(41.92)	(31.74)	(36.95)
IDP	355.73	383.92	263.44	268.05	305.40
	(96.31)	(133.20)	(42.20)	(39.26)	(35.60)

Table D.5

FITCHBURG YEAR 2: PREDICTIONS (STANDARD ERROR)

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	522.62	506.71	576.25	478.55	476.68
	(110.99)	(125.34)	(98.10)	(71.81)	(45.70)
P25	252.02	289.31	375.26	301.36	384.61
	(52.32)	(174.95)	(82.95)	(59.61)	(42.52)
P50	154.99	231.75	277.34	261.45	280.15
	(37.62)	(258.14)	(88.95)	(78.09)	(46.86)
PFD	238.49	167.52	259.04	269.23	319.49
	(50.99)	(190.43)	(60.07)	(61.05)	(44.76)
IDP	565.92	578.76	309.80	292.95	357.59
	(240.52)	(147.01)	(59.86)	(52.59)	(40.62)

Table D.6

FRANKLIN COUNTY YEAR 1: PREDICTIONS (STANDARD ERROR)

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	339.41 (48.85)	331.16 (85.78)	363.88 (42.71)	307.69 (32.14)	373.98 (31.99)
P25	194.03 (36.72)	202.90 (119.14)	233.70 (35.73)	206.40 (28.86)	310.08 (31.79)
P50	790.58 (622.94)	835.98 (194.02)	257.36 (61.36)	223.46 (49.40)	290.51 (43.80)
PFD	204.63 (63.47)	195.91 (115.50)	139.11 (21.11)	149.24 (22.55)	248.11 (28.12)
IDP	304.99 (72.11)	304.31 (101.88)	252.72 (33.73)	253.79 (32.09)	306.57 (30.91)

Table D.7

FRANKLIN COUNTY YEAR 2: PREDICTIONS (STANDARD ERROR)

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	394.27 (79.90)	400.15 (60.24)	368.95 (49.45)	289.23 (32.31)	399.09 (36.82)
P25	216.31 (37.63)	222.84 (84.21)	325.52 (56.53)	251.95 (36.74)	361.40 (38.28)
P50	452.63 (200.38)	500.95 (136.85)	297.26 (80.99)	252.19 (61.45)	282.68 (47.21)
PFD	148.91 (36.46)	145.71 (83.45)	136.19 (24.05)	148.02 (25.25)	264.15 (33.34)
IDP	286.11 (65.29)	261.68 (72.44)	238.93 (36.35)	214.81 (28.62)	318.60 (34.21)

Appendix E
PLAN RELATIVES FROM OTHER SITE-YEARS FOR ALTERNATIVE MODELS

Each plan is presented as a percentage of the free plan for that site and year.

Table E.1

PLAN RELATIVES FOR DAYTON YEAR 3[a]

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	100 (-)	100 (-)	100 (-)	100 (-)	100 (-)
P25	95 (0.21)	90 (0.35)	60 (2.65)	63 (2.93)	72 (3.62)
P50	53 (3.11)	53 (1.53)	39 (4.22)	49 (4.10)	63 (4.14)
PFD	76 (1.34)	80 (0.73)	41 (4.16)	42 (5.08)	62 (5.00)
IDP	140 (0.48)	147 (1.25)	45 (3.39)	55 (3.07)	67 (3.56)

[a]Absolute t-values given in parentheses are based on the difference between that plan and the free plan as measured on the dollar scale, not in proportions.

Table E.2

PLAN RELATIVES FOR SEATTLE YEAR 1[a]

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	100 (-)	100 (-)	100 (-)	100 (-)	100 (-)
P25	89 (0.53)	95 (0.30)	82 (1.35)	83 (1.48)	85 (1.95)
P50	--	--	--	--	--
PFD	68 (2.02)	70 (1.78)	52 (4.01)	57 (4.18)	73 (3.61)
IDP	68 (2.43)	65 (2.18)	62 (3.18)	69 (2.93)	86 (1.82)

[a]Absolute t-values given in parentheses are based on the difference between that plan and the free plan as measured on the dollar scale, not in proportions.

Table E.3

PLAN RELATIVES FOR SEATTLE YEAR 2[a]

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	100 (-)	100 (-)	100 (-)	100 (-)	100 (-)
P25	128 (0.67)	135 (1.07)	75 (1.85)	86 (1.20)	81 (2.37)
P50	--	--	--	--	--
PFD	68 (1.21)	77 (0.67)	52 (3.88)	72 (2.42)	72 (3.57)
IDP	67 (1.40)	66 (1.07)	59 (3.31)	72 (2.65)	81 (2.45)

[a]Absolute t-values given in parentheses are based on the difference between that plan and the free plan as measured on the dollar scale, not in proportions.

Table E.4

PLAN RELATIVES FOR FITCHBURG YEAR 1[a]

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	100 (-)	100 (-)	100 (-)	100 (-)	100 (-)
P25	67 (0.85)	66 (0.82)	81 (1.00)	97 (0.16)	90 (1.08)
P50	43 (1.56)	50 (0.90)	59 (2.03)	62 (2.14)	71 (2.70)
PFD	60 (1.04)	61 (0.88)	49 (3.26)	53 (3.40)	76 (2.44)
IDP	72 (0.70)	80 (0.53)	60 (2.64)	78 (1.52)	81 (1.98)

[a]Absolute t-values given in parentheses are based on the difference between that plan and the free plan as measured on the dollar scale, not in proportions.

Table E.5

PLAN RELATIVES FOR FITCHBURG YEAR 2[a]

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	100 (-)	100 (-)	100 (-)	100 (-)	100 (-)
P25	48 (2.21)	57 (1.00)	65 (1.77)	63 (2.10)	81 (2.07)
P50	30 (3.14)	46 (0.96)	48 (2.44)	55 (2.19)	59 (3.68)
PFD	46 (2.33)	33 (1.48)	45 (3.06)	56 (2.43)	67 (3.16)
IDP	108 (0.16)	114 (0.37)	54 (2.65)	61 (2.33)	75 (2.74)

[a]Absolute t-values given in parentheses are based on the difference between that plan and the free plan as measured on the dollar scale, not in proportions.

Table E.6

PLAN RELATIVES FOR FRANKLIN YEAR 1

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	100 (-)	100 (-)	100 (-)	100 (-)	100 (-)
P25	57 (2.38)	61 (0.87)	64 (2.64)	67 (2.62)	83 (2.13)
P50	233 (0.72)	252 (2.38)	71 (1.54)	73 (1.52)	78 (1.88)
PFD	60 (1.68)	59 (0.94)	38 (5.20)	49 (4.44)	66 (4.34)
IDP	90 (0.40)	92 (0.20)	69 (2.33)	82 (1.34)	82 (2.30)

[a]Absolute t-values given in parentheses are based on the difference between that plan and the free plan as measured on the dollar scale, not in proportions.

Table E.7

PLAN RELATIVES FOR FRANKLIN YEAR 2

Plan	ANOVA	ANOCOVA	One-Part Model	Two-Part Model	Four-Part Model
Free	100 (-)	100 (-)	100 (-)	100 (-)	100 (-)
P25	55 (2.02)	56 (1.71)	88 (0.66)	87 (0.84)	91 (1.05)
P50	115 (0.27)	125 (0.67)	81 (0.81)	87 (0.56)	71 (2.38)
PFD	38 (2.79)	36 (2.46)	37 (4.66)	51 (3.74)	66 (3.79)
IDP	73 (1.05)	65 (1.46)	65 (2.41)	74 (1.89)	80 (2.49)

[a]Absolute t-values given in parentheses are based on the difference between that plan and the free plan as measured on the dollar scale, not in proportions.

[redacted]