A Cross Sectional Analysis of Air Pollution's Impact on Chronic Respiratory Disease in Ontario

by

Christopher Alan Duddek

B.Sc. University of Manitoba, 1989

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

August 1994

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

(Signature)

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date August 13, 1994

DE-6 (2/88)

Abstract

Does ambient air pollution in Canada pose a threat to respiratory health? For a study initiated by Health Canada, we combined analyses of micro-level data from both the 1990 Ontario Health Survey with an environmental air monitoring system to obtain a quantitative answer to the question.

In contrast to studies designed to collect special purpose data, the Ontario Health Survey was not designed to address respiratory health issues. In spite of this, this cross sectional database was rich enough for modelling. We used asthma and emphysema as the response variables in assessing the impact of four pollutants estimated for summer and winter.

Two analyses were conducted for each response variable, one incorporating survey design information the other ignoring it. Age, income, smoker type and sex were significantly related to asthma at a $\alpha = 5\%$ level of confidence in both analyses. None of the pollutant covariates figured in the model.

Using the classical χ^2 test for nested models as the criterion, the emphysema model achieved a better fit than the asthma model. Smoker type and age, in particular, were strongly related to emphysema; income and number of cigarettes smoked were significantly but less strongly related; summer NO_2 was marginally significant, depending on which of the two analyses was considered.

Contents

List of Tables

List of Figures

Acknowledgement

This thesis was made possible by Jim Zidek and Rick Burnett: both were instrumental in bringing the Health Canada project to the University of British Columbia. I appreciate their unending labour in sorting through the red tape associated with such a project.

To get me thinking about my thesis, the biostatistics subgroup met regularly in the summer of 1993. I would like to acknowledge the regulars for their help: Victor Espinosa-Balderas, Nhu Le, John Petkau, Rick White, Hubert Wong, Weimin Sun and Jim Zidek.

At Health Canada, I was lucky to have a good ally in Robert Tkalec. When I asked questions or complained about incomplete documents he was quick to the whip.

Weimin Sun deserves special recognition for his modification and implementation of the spatial interpolation methodology developed by Jim and Nhu. When I needed data he would work until the next morning's sunrise. His dedication to the task was remarkable.

On the social side, I would like to recognize those who made my UBC days exciting. Xiaochun Li was always up for movies and dinner when I was stuck or tired; Nita Deerpalsing was stellar as my perpetual audience, both for the thesis and in general; and Tim Fijal was continually breaking me up with his Hippocrates' saying that 'wholemeal bread clears out the gut'.

Finally, I would like to thank Jim Zidek for his neverending encouragement. Without it I may never have succeeded in finishing the thesis.

To my mom and dad

for producing me.

The yellow fog that rubs its back upon the window-panes,
The yellow smoke that rubs its muzzle on the window-panes
Licked its tongue into the corners of the evening,
Lingered upon the pools that stand in drains,
Let fall upon its back the soot that falls from chimneys,
Slipped by the terrace, made a sudden leap,
And seeing that it was a soft October night,
Curled once about the house, and fell asleep.

⋮

And should I then presume?
And how should I begin?

From *The Love Song of J Alfred Proofrock* by T.S. Elliot

1 Introduction

Resource exploitation was a dominant ideal of European imperialism. Dreams about the 'New World' were predicated on animal fur, vast stands of East Coast forest and unlimited stocks of Grand Banks cod (Seiler, 1993, 303).

Industrialization, urbanization, technology and the rise of bureaucracy, however, radically altered the landscape. Our ability to deleteriously effect our environment gradually led to the emergence of agencies now associated with the welfare state. Health Canada, Agriculture Canada, the Environmental Protection Agency and the Food and Drug Administration are a few of the North American institutions which attest to our faith in resource management.

The air we breathe has come to be seen as one of those resources. Two approaches have been taken to ensure and improve air quality: emission controls and development of ambient air quality standards. The automobile gives a good example of the first. Detroit manufacturers continuously update their line of cars, thereby achieving opportunities to incorporate emission reducing engineering improvements. Since preventative measures are viewed as the least costly of pollution controls, regulatory agencies in the United States have goaded the car industry into meeting higher auto emission standards. By the mid 1990s Canada will have realigned standards so that they fall in line with America's. Because of this vigilance, North American standards are higher than those of many European countries (Hoberg, 1993, 108-109).

Ambient air quality standards provide another benchmark. The need for them became apparent after the London fog episode of 1952. The maximal 24-hour concentration of sulfur dioxide, ten times the current allowable concentration, resulted in an estimated 4000 excess deaths (Griffith, 1989, 112). The ambient air quality standard serves as a benchmark and is supposed to be determined in spirit in accordance with the current state of scientific evidence. With general acceptance of these principles, ambient air quality standards have become a cornerstone of public health policy.

The concept of the Threshold Limit Value or TLV is a key to understanding the history of

air quality guidelines in the States. Since the 1940s TLVs have been set by the TLV Committee of the American Conference of Governmental Industrial Hygienists. The TLV Committee officially defines the TLV to be levels at which "nearly all workers may repeatedly be exposed day after day without adverse health effects." The evidence, however, suggests that the TLV Committee has always been sensitive to the impact their decisions would have on industry. They thus chose levels thought to be achievable by major players in industry (Rappaport, 1993).

The role of the TLV should become clear after sketching air pollution regulation in the U.S. from 1970 to present. The inadequate regulation of hazardous air pollutants was implicitly recognized by the Clean Air Act of 1970. It gave the Environmental Protection Agency (EPA) the authority to impose emission standards that guaranteed "an ample margin of safety." The EPA immediately became sensitized to the potential for economic dislocation if stringent emission limits were set and bypassed enforcing the law by avoiding to list and regulate airborne toxicants. When political pressure over specific substances arose that was too great to ignore, the EPA established emission limits based on economic rather than health considerations (Robinson and Paxman, 1992).

In the early 80s, commencing with the inauguration of Ronald Reagan, the EPA sought to delegate responsibility of airborne pollutant regulation to the states. The states then developed Ambient Air Level guidelines which were based on TLVs multiplied by an appropriate safety factor. But is the TLV a good starting point for air quality guidelines? There are reasons to be wary of the TLV.

First and foremost, the TLV is tainted by the TLV Committee's decision process:

> TLVs for particular substances were heavily influenced by corporations with direct financial interests in the substances being evaluated. . . . TLVs often represent the exposure levels actually prevalent in major firms rather than levels at which no adverse health effects are reported (Robinson and Paxman, 1992, p. 392).

Whose voices are not heard by the Committee? Where economic interests have had a stake, collective health has partially been compromised to minimize the negative fiscal impact of higher emission standards on the offending industries. As long as the hazard is not deadly in the short term and effects only sensitive populations, profits have in past held the upper edge in public health policy.

A second problem with the TLV derives from its use. States use the TLV by multiplying its value by a constant. But how is the constant determined? Since the TLV is based on the 40 hour workweek, multiplying by 4.2 would account for the 168 hour week experienced by residents living in the affected area. Still, the guideline is supposed to be designed for working populations and not necessarily for a population encompassing the more sensitive segments like the very young and very old. The resulting state defined Acceptable Ambient Air Level guideline for acrylonitrile, for example, varies by a ratio of over a thousand for regulating states. The variability alone casts suspicion on the process.

The Clean Air Act Amendment of 1990 repositions the EPA as the national standard bearer in air pollution policy. The Act requires the EPA to develop national emission standards for 189 toxic substances in the next decade. In contrast to the "ample margin of safety" quoted in the Clean Air Act of 1970, the Amendment requires that standards be set on "maximum achievable control technology." This reorientation is proposed to accelerate the pace of standard setting by basing it on technological advancement. The TLV, as a result, will play a lesser role in health policy (Robinson and Paxman, 1992).

Gibson (1989) documents a similar tale of corporate influence over public air pollution policy in Ontario. The Inco smelter in Sudbury has long been an infamous source of $S0_2$ emissions. Little headway in reductions have been made, however, because Ontario Environment Ministry officials knew "that the environmental benefits would be less immediate and perceptible than the costs of abatement and that the beneficiaries would be more dispersed and less well connected politically than the recipients of abatement orders" (Gibson, 250).

Ignoring the decision making realities of the political realm, the question remains: how seriously is human health affected by air pollution? A call for quantification, in the form of environmental health impact studies, is a common response (Britton, 1992), with many forms having been attempted. I give a flavour of recent work in the literature review. A taxonomy of epidemiological studies is provided in the chapter covering cross sectional designs.

In a bizarre twist, regulatory procedures often provide data for the studies. Once an air quality standard is determined, pollutant monitoring stations are erected and data collected. Thus, ironically, the available data on pollutant exposure is driven by regulation rather than for its utility for inquiries into public health (Matanoski et al., 1992).

In this study we evaluate the risks of air pollution to chronic lung function. The 1990 Ontario Health Survey provides us with a cross sectional view of the health status and socioeconomic level of Ontario's population. The air pollution data arises from an air monitoring network of 37 stations, tracking four pollutants, from 1983 to 1989. The study's methodology and analysis are given in greater detail in the appropriate chapters.

The objective of this study is ambitious. As opposed to chronic effects the study of acute effects, as measured by longitudinal hospital admissions, is more common. The difficulty with a study of chronic effects is the measurement of exposure. Long lag times, differences in pollutant mixtures over time, movement of individuals in the study population and error in self reported medical conditions are but a few of the obstacles to valid inference.

On the bright side, the Ontario Health Survey data is comprehensive and contains information on a large sample. The six years pollution data employs recently developed methodology: where necessary the multivariate air monitoring station readings are spatially interpolated (Brown, Le and Zidek, 1993).

To make informed policy decisions on the effect of pollution on respiratory health we need to employ pre-existing data sources wherever possible. This study provides that kind of opportunity.

2 Literature Review

The literature on environmental health impact assessments is voluminous. In this chapter I cover some of the more recent papers in this area. I have ordered the material to follow a natural flow. Since experimental studies can more or less stand on their own I discuss them first. Observational studies tend to be more complex. I have apportioned a section to the difficult area of exposure measurement and one to cross sectional studies of a similar intent.

2.1 Experimental Studies

Although questions about a pollutant's effect on human populations can be ethically difficult to address in an experimental study, some have been carried out. Whenever an effusion of possible predictors with many levels appears, experimental design, R.A. Fisher's territory, tempts good researchers. In this section I describe recent experimental studies.

Hackney et al. (1992) evaluate the acute effects of nitrogen dioxide (NO_2) on older adults with chronic obstructive pulmonary disease. They note that although animal toxicological studies prove high doses of NO_2 pose a respiratory health risk, the epidemiological literature is inconclusive. They increase the power of their study by focusing on a sensitive segment of the population. This strategy has two benefits. First, the sensitive subpopulation is of interest in its own right since, by definition almost, they are most affected by pollutant levels. Second, information garnered from sensitive subpopulations may have implications for the larger population. We could use the metaphor of individuals as instruments: sensitive individuals might be like us except that they react more strongly to the same stimulus.

Hackney's Los Angeles study combines laboratory and field work. In the laboratory researchers examined the lung dysfunction of subjects exposed to 0.3 ppm NO_2 for four hours interlaced with four bouts of exercise. In the field they evaluated the exposure measurements from personal exposure monitoring devices worn by the 26 volunteers for two week periods. NO_2 readings are traditionally highest in LA during fall and winter. The study interval was

no exception with an average of 125 $\mu g/m^3$ as compared to the annual average of 90 $\mu g/m^3$.

Two conclusions are drawn. First, from a comparison of station to personal exposure measurements, NO_2 exposure is strongly influenced by outdoor pollution, even though up to 90% of the subject's time is spent indoors. Second, the sensitive population's short term NO_2 exposure results in little short term clinical exacerbation of respiratory disease.

McDonnell et al. (1991) study the effect of a 6.6 hour ozone (O_3) exposure on 38 healthy humans. The response measure, forced expiratory volume in one second (FEV_1), shows significant decline upon comparing clean air exposure to 0.08 ppm O_3.

McDonnell, Muller, Bromburg and Shy (1993) improve on the previous study by examining more design points, increasing the range to 0.0–0.4 ppm O_3 and increasing the sample size. This time they divide the sample into an exploratory sample of 96 subjects and a confirmatory sample of 194 subjects. The first sample specifies a model after fitting many models and predictors showing statistical significance. The second sample validates the model and protects against declaring predictors significant when they spuriously appear in the model by chance alone. The study concludes that O_3 explains 31% of the response variance while age explains 4%.

Experimental studies necessarily examine acute effects. From the three I reviewed, ozone seems to be more potent than nitrogen dioxide. Though experimental studies are useful for corroborating evidence produced by observational studies, the artificial context of a laboratory chamber may not correspond to pollution effects occurring in daily life. Moreover we cannot assess chronic health effects. Observational studies address these issues head on. The first step in an observational study is to estimate the extent of exposure.

2.2 Estimating Cumulative Ambient Air Pollution Exposure

In observational studies we must deal with the problem of quantifying internal dose of the agent over time. Difficulties like the impossibility of obtaining internal dose data, working with imperfect proxies and refining pollution exposure modeling strategies characterize the

quest for an adequate solution.

How can we use ambient air pollution measurements as a surrogate for internal dose? Typically the data come from fixed site air monitoring stations, stations for which data has already been collected over long periods of time. The objective is to link station data to human populations. Commonly the data are first interpolated spatially to other sites where no measuring equipment exists. Then the exposure based on living patterns reported by individuals in a survey are modelled.

Abbey, Moore, Petersen and Beeson (1991) address the interpolation question. They validated a simple method of interpolation by fixed site monitoring station deletion: after a station was deleted, interpolated values were calculated and compared against the measured values.

Ozone (O_3) and total suspended particulate (TSP) were measured for at least three years by 126 and 142 stations, respectively. Ozone was measured for one minute every hour and TSP for 24 hours every sixth day.

The statistics of interest were exceedance frequency and mean concentration. The exceedance frequency was compared to US regulatory policy levels which are often stated in terms of maximum allowable values.

Their interpolation method obtained estimates for all ZIP codes in California. A maze of rules were put together to get interpolated values. First, a measuring station was valid if at least three years of pollution data existed. Second, for a given ZIP centroid, a station was valid if it resided within a radius of 50 km. Third, a series of three zones, based on concentric rings, were defined, zone A being the closest, zone B in between and zone C the furthest. The only stations used in the interpolation were those falling within the closest ring. Up to three stations were used in a given zone.

The zone definitions differed for each pollutant. TSP was assumed less homogeneous over space and the radius boundaries for the zones were tighter than for ozone. In the study area about 60% of the population lived within 10 km of a TSP monitoring site (zone B or better).

About 90% lived within 16 km of an ozone monitoring site (zone A or better).

The study concludes that the interpolation methods worked well in their particular case. The correlation of 0.78 for TSP and 0.87 for ozone suggest the importance of treating different pollutants differently since, even with tighter controls of TSP zones and a greater number of stations, the TSP estimates were more inaccurate than those for ozone. Second, if persons can be situated to ZIP code centroids for significant time periods, the interpolation method may produce estimates which are good surrogates for internal dose.

Seixas, Robins and Becker (1993) propose to model human exposure using the occupational history of individuals and hundreds of thousands of occupational ambient air pollution measurements. They motivate their research by contending that simple exposure estimates do not adequately capture the complexity of the disease process. The implicit assumptions that dose is a linear function of concentration and independent of time go against evidence of nonlinear toxicological behaviour of many substances causing adverse chronic health effects. In addition, the use of a simple statistic for exposure contributes to error in variable bias.

Their modeling solution estimates the exponents of pollutant concentration and time between measurement and reported outcome. In their application 300,000 dust exposure observations from the shafts of underground coal miners were used to estimate model parameters. Each of the 1200 respondents provided work histories sufficient to obtain estimates of individual exposure. Despite the efforts made, the predictive power of alternative, simpler models achieved competitive performance levels.

In conclusion, for fixed site air station monitoring data, the best epidemiological studies can do to approximate individual internal dose over time is to use some combination of interpolation and human exposure modeling. When stations are close enough, e.g. within 10 km for TSP or 16 km for ozone, a crude interpolation method will do well. For greater distances the quality of the interpolation will decrease, though the decrease is pollution specific and the degree of quality deterioration remains to be further quantified.

The subtlety of human exposure modeling promises to challenge researchers striving for the ideal. Models, however, will continue to be data dependent: the availability of work histories, migration patterns and activity diaries will determine the quality of inference.

More research is needed to answer the difficult question of how internal dose relates to exposure measures. If progress is made, we will also be able to evaluate the soundness of the modeling approach. Until then, modeling will offer hope of improving the power of epidemiological studies.

2.3 Cross Sectional Studies

I will characterize five recent cross sectional studies. Their variety of approaches is striking. While some of the studies simply analyse differences between study and control groups, others use standard regression techniques. The sample sizes range from 600 to 3900 respondents. The findings also show a range. Causes of respiratory symptoms go from grain-farming to blowing alkali salts.

Abbey, Moore, Petersen and Beeson (1993) used Seventh-day Adventist nonsmokers to check on TSP, ozone and sulfur dioxide. Their logistic regression analysis found significant relationships between ambient concentrations of TSP and ozone with several respiratory disease outcomes.

Senthilselvan, Chen and Dosman (1993) examine the relationship between grain farming and respiratory illness in Humboldt, Saskatchewan. They split their study population into four cohorts, use a survey to obtain binary responses and compare prevalences between cohorts. Asthma was found to be significantly related to grain farming and sex; wheezing was related to grain farming and smoking.

Gomez, Parker, Dosman and McDuffie (1992) considered the effect of alkali dust on a Southern Saskatchewan population living near Old Wives Lake. When prevalences were compared their control and study groups' chronic wheeze, eye irritation and nasal irritation increased. Unlike the previous two studies, the researchers used forced vital capacity (FVC) and forced

expiratory volume in one second ($FEV_{1.0}$) in their analysis.

Xiping, Dockery and Wang (1991) measured the FVC and $FEV_{1.0}$ from a sample of Beijing residents. Besides underlining the importance of coal heating as a source of respiratory problems they discovered a relationship between outdoor SO_2 and FVC and $FEV_{1.0}$ in subjects who had not used stove coal heating.

Özkaynak and Thurston (1987) use U.S. mortality statistics to evaluate the effect of ambient air pollutants. Their regression model suggests that SO_4 concentration is a significant factor in mortality prediction.

The published cross sectional studies provide some evidence that respiratory health is adversely effected by irritants in the air. This study will add to the continually growing collection.

3 Issues in Cross Sectional Analyses

Before analysing the data I will describe, in broad terms, epidemiology, and then focus in on the virtues and problems of cross sectional analyses.

As a science, epidemiology has had a number of high impact successes. I document the case of AIDS as a way of illustrating the role and importance of epidemiologists. Historically, the etiology of infectious disease has been easier to discover than that of noninfectious disease. Noninfectious disease is more of a problem in the industrialized world, however. With an aging population it would not be far-fetched to suggest that the quality and length of life will in part be determined by our ability to understand and control noninfectious disease.

Noninfectious disease etiology is a slippery concept. With multiple causes and long developmental intervals, epidemiologists are forced to consider sophisticated analytic techniques. As a result, the literature is imbued with a variety of study methodologies. I provide a short taxonomy of epidemiological studies as an introduction to this area. A more detailed look at the cross sectional study ends the chapter.

3.1 Definition of Epidemiology

Epidēmos is Greek for prevalence. Hippocrates wrote several books concerning disease prevalence in the fourth century BC. In one, he distinguishes between endemic diseases, which prevail continuously at relatively low levels, and epidemic diseases, which occur at higher than expected frequencies. When his writings focus on the health of populations rather than individuals Hippocrates is donning the epidemiologist's hat.

Epidemiology, as practiced today, "is the study of the distribution of states of human health and of determinants of deviations from health in human populations" (Valanis, 1992). This report is an epidemiological study by virtue of its concern with distribution of chronic respiratory illness over a geographic area and its relationship to airborne pollutants.

3.2 The Disease Process

Disease is a complex process that can be separated into stages of prepathogenesis and pathogenesis. Prepathogenesis refers to the initial bodily changes that may or may not lead to pathogenesis. The individual's exposure to one or more agents comprises the first stage of prepathogenesis. If the individual is susceptible he or she is unable to adapt to introduction of the agent; with successful adaptation the disease does not develop any further.

Pathogenesis occurs when the disease has successfully established itself within the host. During early prepathogenesis, events take place that are clinically difficult to detect. This period is often referred to as the presymptomatic or preclinical stage. At latter stages, however, clinical symptoms show up. This is the point where the person is commonly said to 'have' the disease. Identification and classification of disease is complicated by partial symptoms, misdiagnosis and long lag times between early prepathogenesis and development of clinical symptoms.

Three components are distinguishable in the disease process (Valanis, 1992, Chapter 2). The first, denoted as the *host*, is the site of the disease. Factors relating to the host's susceptibility, like lack of sleep, malnutrition, aging and immunity, may be useful in understanding the transition from healthiness to pathogenesis and the rate of that transmission. Explicit descriptions of the host, e.g. the human subject, is a good first step in describing disease.

The *agent*, initiator of the disease process, is the second component of the trinity. The *tubercle bacillus*, without which tuberculosis would be unknown, is an example of a parasitic agent. Risk factor assessments try to describe and quantify the hazards agents pose. Most infectious diseases arise from a single agent; noninfectious disease are usually the result of multiple agents.

The *environment* is the third component and relates host to agent. The environment is the sum of external conditions and influences affecting the life of living things. By this definition, physical, biological and socioeconomic descriptors are encompassed. The environment can

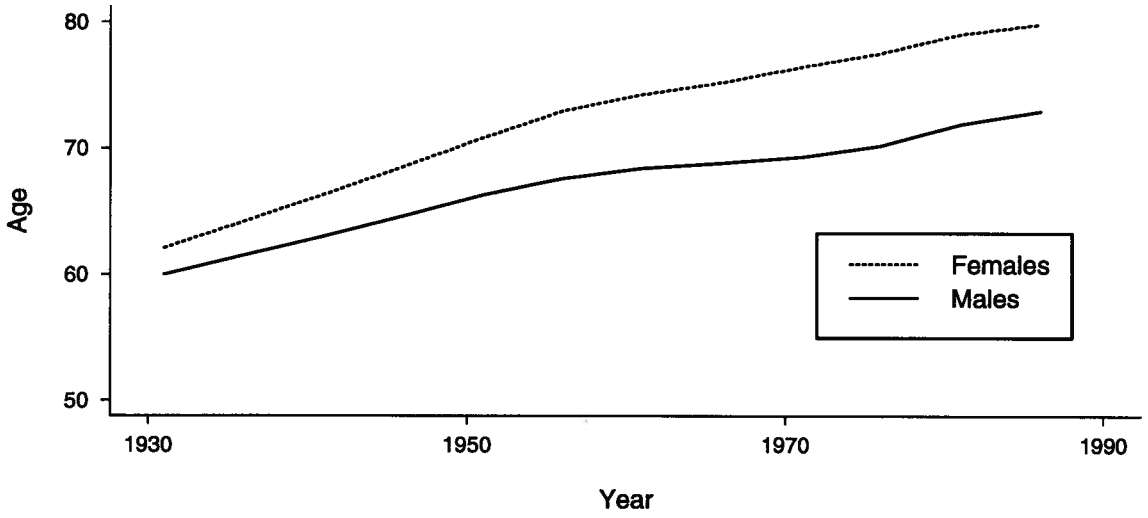

Figure 1: Canadian life expectancy at birth.

effect host susceptibility and viability of the agent. Cholera, for example, spreads best when people live in crowded conditions with poor sanitation. In this study wind patterns and temperature may have a significant effect on an individual's exposure to pollutants.

Thus, disease is a rather complicated process. The focus of epidemiology, as opposed to practitioners of clinical medicine, say, is on the triangular relationship between host, agent and environment. To uncover the etiology or causes of disease is the goal of risk assessment studies.

3.3 Trends in Disease Patterns Since 1900

Life expectancy in Canada, as shown in Figure 1 (Sources: Statistics Canada, 1983, and Institute for Health Care, 1990), has increased steadily since the early part of the twentieth century. The primary reasons for the increase lie in improved sanitation, better diet and our increased ability to control infectious disease. They are interrelated, of course, since a better diet reduces susceptibility and improved sanitation diminishes the opportunity for the spread of virulent bacteria. The strides made in medicine's attack on infectious disease, however, has been nothing less than staggering. In terms of mortality about 45% of deaths in

1900 were attributable to infectious disease. The corresponding figure in 1987 was about 5% (Valanis, 1992, p. 28).

Epidemiology has had some of its greatest successes with infectious diseases. Take, for example, the recent role played by the Centers for Disease Control (CDC) in Atlanta in identifying aquired immunodeficiency syndrome. CDC happened to be the sole supplier of pentamidine, an experimental drug used in chemotherapy and radiotherapy to treat cancer patients. When physicians in Los Angeles and New York City greatly increased demand for pentamidine, epidemiologists at CDC took note. In 1981 CDC's *Morbidity and Mortality Weekly Report* contained two articles about a new disorder termed aquired immunodeficiency syndrome (AIDS).

The new disease had the marking of an infectious disease: AIDS spread among people who had been in contact with one another. Assuming that AIDS was caused by an unidentified virus, a series of risk assessment studies uncovered the connection that the vast majority of AIDS patients were either drug users or homosexuals. The finding had major policy implications. Public education programs commenced, health care resources were reallocated and AIDS research emerged. In short the new disease along with associated risk factors was identified early enough to allow public institutions to reformulate policy and adjust to new realities.

The assumption of AIDS' infectious nature turned out to be right. The human immunodeficiency virus (HIV) was discovered by a French team in 1983. They completed the cycle from initial recognition to a reasonably complete etiology of the disease (Purtilo and Purtilo, 1989). Significantly, these investigations focused on a disease that had a single cause, developed relatively quickly and resulted in a sharp increase in the number of observable cases.

Unfortunately for epidemiology's rising star, noninfectious diseases remain the leading causes of death. They tend to be harder to identify, take longer to develop, have multiple causes and come with long lag times between introduction of the agents and development of definitive clinical symptoms. In the wake of these developments, epidemiology must turn to

more sophisticated methods and measurements to get at the complex etiology of noninfectious disease.

3.4 Types of Epidemiological Studies

In this section I contextualize the cross sectional study by describing its placement within a multitude of evaluative methods. Much of my insight derives from Valanis' book (1992, chapter 3).

The study of disease etiology generally proceeds in an orderly manner from generating hypotheses to determining the causal mechanisms underlying the observed phenomenon. Three types of studies, sequentially related and characterized by increased effort, higher costs and greater investigative controls, are used in epidemiology: descriptive, analytic and experimental.

The descriptive study relies on easily accessible data, often mortality rates, which is analysed in a simple manner. Analyses may consist of breaking the population down by certain characteristics and then comparing mortality rate differentials. Descriptive studies are useful for uncovering unusual events or problems with data quality and can be thought of as equivalent to initial data analysis.

The analytic study relates disease to agents while attempting to control for potentially confounding covariates. At least three types of investigation fall within this category: cross sectional, case control and cohort studies. The analytic study generates hypotheses, leads to experimental studies, and tests hypotheses in an attempt to explain phenomena arising in descriptive epidemiology.

The experimental or designed study distinguishes itself from observational studies, i.e. descriptive or analytic, in terms of control. In an experimental study the researcher selects factors of potential importance, their levels of application and randomly assigns experimental units to the prespecified treatments. After controlling to whatever extent possible for external conditions, the researcher observes the outcomes to determine the importance of the selected factors. With an observational study, the researcher has no control over factor levels or the

assignment of experimental units.

The experimental study seeks to confirm or disavow certain cause and effect relationships. The investigator incorporates experimental randomization to avoid the pitfalls of systematic bias introduced by human intentionality. Under realistic conditions, only people who have the disease of interest are included in experimental studies. For one group, the experimentor removes the suspected agent from the environment; the other group acts as a control. The two groups are then followed over time to see if significantly different changes occur as a result of the treatment.

Since this study is analytic rather than experimental, each of the three analytic subtypes, namely cross sectional, case control and cohort, will be described in greater detail. The cross sectional study is like a snapshot: population characteristics are conceptually sited at a single time point. The resulting measurement of disease prevalence explains why some authors use the term 'prevalence study.'

The case control and cohort studies differ from the cross sectional in that they follow individuals over time. Case control studies begin with two distinct populations: people with the disease and people without. Disease relationships are determined by examining both groups' exposure to a variety of agents and uncovering the greatest differences between the two. In other words, case control studies work from disease to uncover exposure status.

As opposed to case control studies, cohort studies start from exposure status. The best example is the Framingham study begun in 1949 and continued to the present. A random sample of individuals was drawn from the population of Framingham, Massachusetts. Physicians determined that approximately 5000 of the sampled individuals were free of coronary heart disease thus making them suitable for the study. Every year since, the subjects have gone through a physical examination to (a) assess the health status of their hearts and (b) measure exposure risks. The hope is that the risk factors of those who develop heart disease can be identified. The cohort study can either be historical or prospective in nature.

The case of smoking and lung cancer, as documented from a historical perspective by Clemmesen (1993), is a good example of the potential and pitfalls of descriptive and analytic studies. At the turn of the century, incorrect diagnoses prevented the scientific community from identifying smoking as a problem. Much of the evidence against tobacco was anecdotal. With improvements in identification, some studies uncovered a relationship between smoking and lung cancer. Criticism aimed at the studies, however, mostly concerned with potential confounding factors and poor data quality, undermined their impact.

The development of cancer registries in Mecklenburg and New York in the early 1940s acknowledged concerns over data deficiencies. Five studies, published in 1950, produced common findings on the smoking habits of patients with lung cancer. This confirmation of earlier suspicions eventually led to the first International Symposium on Lung Cancer Endemiology at Louvain in 1952. In 1959, about 100 years after cigarettes had first been manufactured in the U.S., the U.S. Public Health Service officially pronounced their "deep concern" over the increase in age adjusted incidence of lung cancer deaths from 4 per 100,000 in 1930 to 31 in 1956. The long time interval needed to identify the association of lung cancer with cigarette smoking points to the importance of illness classification methods, quality data and confirmatory studies.

3.5 Pros and Cons of the Cross Sectional Analysis

Cross sectional analyses provide information about some aspects of the underlying disease process; other aspects of the process remain hidden. The information primarily comes from a statistical model fitted to the cross sectional data. In this section I will describe what kinds of insight can be gained from the model. I first look at the advantages of the cross sectional study.

The model is intimately related to the data. If the data is a sample from some human population then the inferences about the model are applicable to the sampled population. In other words the study and target population are similar if not the same. Observational studies

are better than experimental studies in this respect. In experimental studies the question of how the study population relates to the target population usually looms unanswered.

In a related way, the subject's exposure level in an observational study is realistic, though hard to measure. With controlled experiments the levels at which pollution concentrations are set often bear no relation to levels experienced by the population of real interest. The design for LD50 experiments, where the objective is to get an estimate of the dose required to kill 50% of the population, is an example.

On a practical level, observational studies give information which cannot be duplicated by experimental studies. Whenever irreparable damage to the observational subject is possible, mice, not men, will be sacrificed. Even the *use* of data from the NAZI hypothermia trials, which were ethically indefensible, is controversial. Observational studies, on the other hand, are nonintrusive.

Another practical consideration is cost. Cross sectional studies often attempt to get information about previous events through a questionnaire. This is one way of evaluating long term exposures of human populations to a variety of potentially toxic chemical compounds. By contrast, designed experiments attempt to manipulate factors so that the effects can be observed. Designed experiments are very expensive when extended over many years and come with the danger that the question under study will lose relevance over time. In this sense, cross sectional studies are efficient.

For cross sectional studies in particular, a model allows the investigator to examine the relationship between response variables and predictors. The model under consideration can vary considerably; explanatory variables that are nominal, ordinal and continuous with non-conforming scales can be handled at the same time. If the standard errors of model coefficients are also estimated, we can assess the importance of the predictors relative to one another. In fact the estimation of coefficients and their standard errors is the foundation upon which the house of cross sectional analyses are built. Cross sectional studies, then, can provide useful

information using standard statistical methodology.

Now I proceed to indicate a few areas in which cross sectional analyses are weak. Having information at one point in time rather than at several is one of them. First and foremost, cause cannot be ascertained since event sequence is unknown. For instance, if fitness level is found to be negatively related to a persistent cough, is the individual's inactivity partially the cause of the cough or is the cough a precursor to lessened activity? Without knowledge of specific events ordered in time, it is difficult to make conclusive statements.

Related to this aspect of cross sectional studies is the type of disease statistic adopted. Epidemiologists make a distinction between incidence, the number of cases of a disease in a prescribed time interval, and prevalence, the proportion of people with the disease. Cross sectional studies measure prevalence, a concern since bias is associated with prevalence: people die or move in response to the occurrence of a disease.

Exposure is one of the key measures in air pollution risk assessment studies. Exposure occurs when the host comes in contact with an agent in the environment. Depending on the exposure data, cross sectional studies can either be ecological or relational. Ecological studies use station pollution data which is assumed to apply to the individuals who live nearby. Relational studies use exposure information collected for every individual selected in the study.

The ecological fallacy arises if one assumes an estimate based on an average applied to individuals is equivalent to an estimate based on individual measurements. If, for example, the distribution of a pollutant is heterogeneous then the absorbed dose among individuals will probably not be the same. Another possibility is that individuals attain different exposure levels due to daily commutes from one area to the next. Thus, a false relationship between pollutants and disease can be observed because the averaged exposure data does not adequately reflect the relation between subjects' absorbed doses in different geographic areas. The ecological fallacy can be seen as arising from measurement error. The sensitivity of regression coefficients to various levels of spatial aggregation has not been studied comprehensively

(Evans et al., 1984).

Another drawback of the cross sectional design arises from the lack of control. Alas, people are not randomly distributed experimental units. They are intentional agents often making decisions related to the area of scientific interest. An example is provided by asthmatics who, aware of their own hypersensitivities, exercise self selection in terms of where they live and where they work (Lebowitz, 1991). Individuals who make informed decisions of that sort should no longer be considered, strictly speaking, observations from a stochastic process that assumes independence.

The context of these studies is important; they ought to be thought about as one part of an extensive, ongoing research effort. By itself, certainly, a cross sectional study cannot prove a causal biological relationship. Bates (1992) suggests we look for coherence in complex phenomena in building a scientific case. Epidemiological evidence should be corroborated, for example, by toxicological studies and results from molecular genetics.

4 Description of the Ontario Health Survey

In Canada, 10% of the 1991 gross domestic product was spent on health care (Evans, 1993, 32). Concurrently, little is known about the health status of the general population. This creates difficulties for provincial governments which must provide health care to individuals at local levels. What data sources are available for these agencies? Are they adequate to meet the need of the increasingly difficult task of optimizing the distribution of federally allocated funds?

The information we do have comes from administrative sources like hospital admission data. Although this data sheds light on who has received medical attention it remains silent about those who do not feed into the system. Further, profile information such as smoking history and socioeconomic status is limited or nonexistent from these sources. The need for better health data is clear.

In 1978-79, Statistics Canada conducted the Canada Health Survey, the first comprehensive health survey taken in Ontario. Successive health related surveys include the Canada Fitness Survey (1981), the Canada Health and Disability Survey (1983/84) and the Health and Activity Limitation Survey (1986/87). For provincial planning purposes, however, the available data had been inadequate. The sample sizes were not large enough to permit inferences below the provincial level. The Ontario Ministry of Health recognized the value of funding a survey which would allow for estimates at the level of District Health Council or Public Health Unit and in 1987 proposed a health survey for the population of Ontario. Four years later the proposal became reality: the survey was carried out and a rich source of new information about the health status of the population of Ontario was created. The database obtained from the 1991 Ontario Health Survey (OHS) provides the information for our study. The large sample size, wide geographic coverage and detailed respondent information will enhance the quality of the study. This chapter gives an overview of the survey.

4.1 Objectives

The survey set out to provide baseline statistical data on the health of the Ontario population at the Public Health Unit level. The objectives are to:

- ▷ measure the health status of the population;
- ▷ collect risk factor data for the major causes of morbidity and mortality;
- ▷ collect data related to socioeconomic and demographic variations in health;
- ▷ measure awareness of high risk behavior;
- ▷ measure utilization of health services;
- ▷ collect descriptive data for health units; and
- ▷ collect data comparable to that in the Canada and Québec Health Surveys.

The long length of the resulting questionnaire reflects a vigorous attempt to achieve all the objectives. The high response burden was recognized from the start and evaluated during pretesting.

The high level of geographical coverage is worth highlighting. As Figure 2 indicates, the Province of Ontario can be divided into thirty seven Public Health Units (PHUs) or districts. The PHU is similar to Census Division, the difference being marginal disagreements in boundaries. Some PHUs, for example, are aggregates of two Census Divisions. During analysis, the PHU will play a vital role in linking geographical pollution data to individuals in the sample. The large size of the PHU ensures that people living within one can reasonably be expected to be bounded to the area in terms of daily movement. We hope there are enough PHUs to enable a good description of pollution differentiation between areas of the province.

4.2 Data Collection Method

We will often refer to the data collection method and so describe it now in some detail. Figure 3 depicts a flowchart summary. I will comment on some of the pros and cons associated with

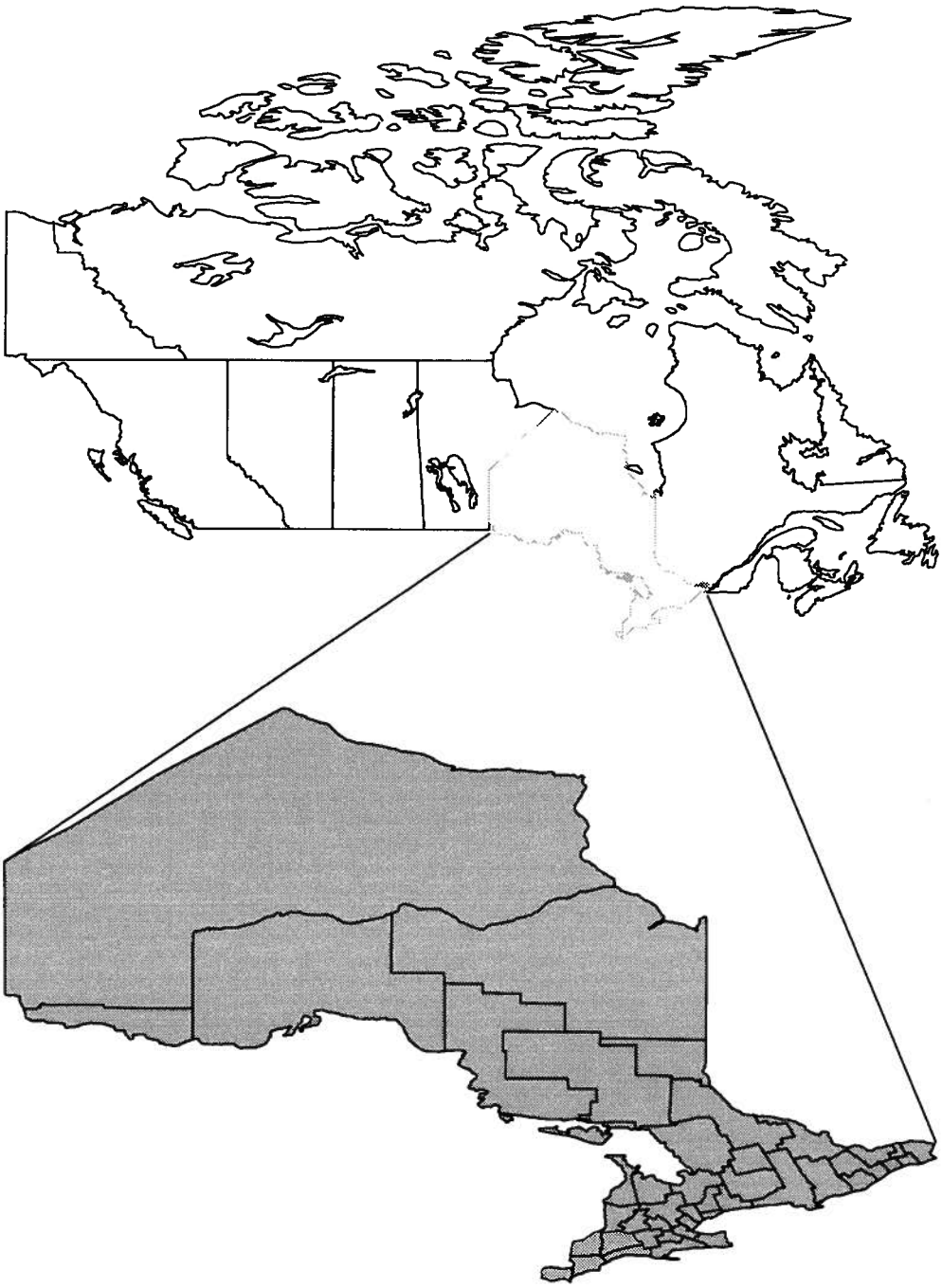

Figure 2: Geographic coverage of the Ontario Health Survey.

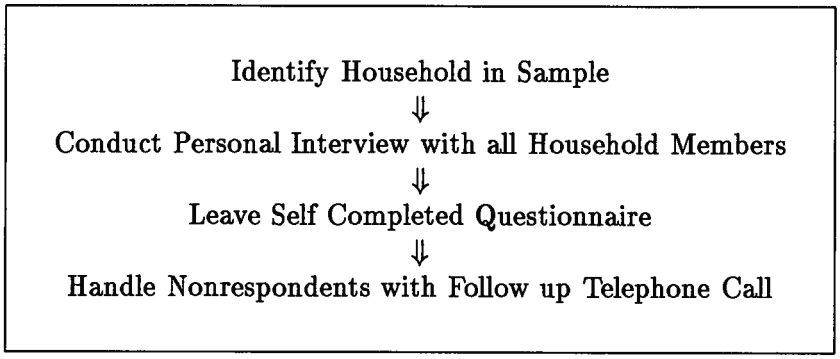

Figure 3: Flowchart summary of the OHS data collection procedure.

each part of the process.

Once a household is identified in sample, a household record form is created. Newly constructed buildings or subdivisions, outdated maps and dangerous neighborhoods can make identification difficult.

As soon as the household record form is available, the personal interview of one household member is possible. That member must be knowledgeable since he reports on all of the other members of the household. Obstacles to successful completion of this phase include inability to contact anyone and errors associated with inaccurate determination of household members. There are at least three benefits from using a personal interview. First, the survey taker gets a chance to introduce the survey without being ignored. Second, empirical observation proves that a higher response rate is generally achieved by a personal interview over self enumeration alone. Third, telephone follow up will be easier since the caller will be making a 'warm' call.

A self completed questionnaire is left for each household member aged 12 and up. If that member does not return the questionnaire before a specified time, two telephone calls are made.

The combination of personal interview, self enumeration and telephone follow up represents a compromise between total survey cost and response rate.

4.3 Target Population

The target population for the interviewer completed portion of the survey is all residents of private dwellings in Ontario during the survey period (January to December of 1990). The 1991 estimate for the population of Ontario is 8.1 million people. As in many surveys conducted by governmental agencies, residents of Indian reservations, inmates of institutions, foreign service personnel and residents of remote areas were excluded.

The target population for the self completed portion of the survey is similar except that the population includes only people aged twelve and up.

4.4 Pretesting

The Ontario Ministry of Health hired Statistics Canada to conduct a pilot survey for the OHS. The four objectives were to:

- ▷ identify weaknesses of the content, wording and structure of the questionnaire;
- ▷ evaluate the efficacy of the training procedures and field operations;
- ▷ quantify the effect of questionnaire length to response rates; and
- ▷ assess the use of an incentive to boost response rates.

In May 1989, a total of 800 dwellings in Peterborough County and the Municipality of Hamilton-Wentworth were surveyed. All households went through the same interviewer portion; four versions of the self completed questionnaire were equally divided among dwellings. Two follow up telephone calls were made to jog the memories of individuals who had not yet returned their questionnaires.

Let {Basic} be the core of the questionnaire (the questions take about 15 minutes to fill out); {Food Frequency Schedule}, the set of questions pertaining to the amount and frequency of different kinds of food the respondent has consumed; {Linkage Information} the set of questions asking for additional identifying information (middle names, maiden names, perviously

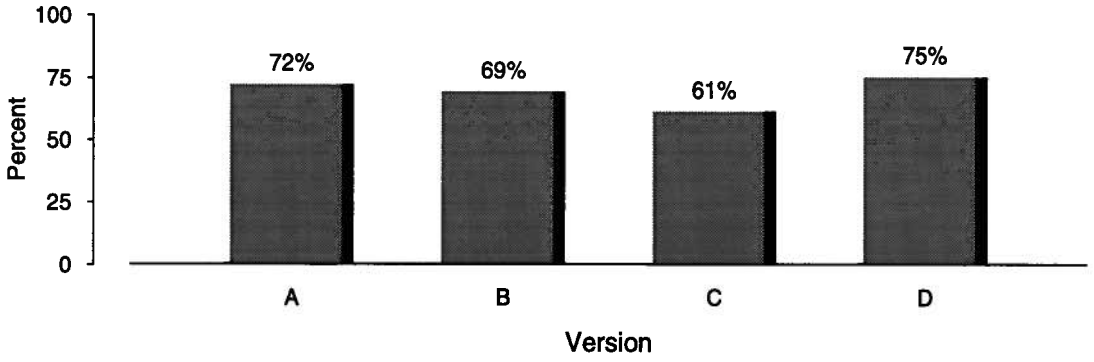

Figure 4: OHS pretest questionnaire response rates.

used surnames and birthplace). The four versions of the self completed questionnaire can then be summarized as follows:

Version A: {Basic}
Version B: {Basic} + {Food Frequency Schedule}
Version C: {Basic} + {Food Frequency Schedule} + {Linkage Information}
Version D: {Basic} + {Food Frequency Schedule} + {Linkage Information} with an incentive.

The incentive in Version D was three $500 prizes drawn at random from those respondents who replied promptly. Two of the objectives were met by including these four versions of the questionnaire.

The response rate for the personal interviewer completed portion was 87%. The overall response rate for the self completed portion was 69%, somewhat under the 75% response rate the Ontario Ministry of Health was shooting for. The main conclusions of the pretest follow from the response rates for each version of the questionnaire, shown in Figure 4. First, adding the food frequency schedule reduced the response rate marginally. Second, requesting the extra identifying information seemed to have a significant adverse effect on the rate. Third, the incentive worked dramatically. The male response rate was 10% lower than that for females. The incentive remarkably increased response rates for males between the ages of 16 to 60. Finally, the use of telephone follow up proved worthwhile since, before calling started, the response rate was below 50%. Although Statistics Canada recommended version A or

version D on the basis of the rate set out by the Ontario Ministry of Health, version B was eventually chosen as the best of all candidates.

4.5 Questionnaire Content

The questions and format of the OHS came from various sources. Previous survey questionnaires like the Canada Health Survey, Québec Health Survey, General Social Survey and Health Promotion Survey were used as models. Potential users of the data, ie. units within the Ontario Ministry of Health such as the Public Health Branch, Public Health Units and District Health Councils, were given an opportunity to develop survey content. Finally, organizations like Statistics Canada and Santé Québec were consulted along the way.

The final form of the questionnaire breaks down into three separate parts: the household record form, the interviewer questionnaire and the self administered questionnaire. The household record form keeps track of the dwelling and the identities of the household members. The content of the interviewer and self administered questionnaires is summarized in Table 1.

The important sections for our study are highlighted with an asterisk in Table 1. For the interviewer completed portion the chronic health problems section contains the two response variables considered in our study, chronic cough and asthma. The personal interview achieves the highest response rate among survey delivery techniques and favorably affects the quality of the response variable data. The other variables, on the other hand, come from the self completed part of the survey.

Sociodemographic data often varies with health outcome. We will therefore want to include a selection of sociodemographic variables from the list. The information derived from questions on the OHS ranges from country of birth to education, income and housing. Comparability of sociodemographic data to other sources, e.g. Canada Census, provides a possible data integrity check.

Information on the multifaceted phenomenon of smoking is extremely important for any study of respiratory health. The section devoted to smoking is comprehensive, identifying

• Household Record Form
• Interviewer Completed Questionnaire
Contacts with Health Professionals
Disability within the last Two Weeks
Use of Medication
Medical Insurance
Accidents and Injuries
Health Status
Restriction of Activities
Chronic Health Problems*
Health Problem Probes
Socio-economic Information*
• Self Completed questionnaire
Your Health
Medicine and Drugs
Smoking*
Alcohol*
Your Family*
Dental Health
Your Life in General*
Driving and Safety
Women's Health
Sexual Health
Occupational Health*
Physical Activities*
Nutrition

Table 1: OHS questionnaire content.

smoker type, the number of cigarettes smoked daily and the ages at which smoking began and ended, where appropriate. As a bonus, questions aimed at the extent of second hand smoke were also asked.

The Short Michigan Alcohol Screening Test (SMAST) score is included in this study even though, on the surface, it may be of peripheral interest. The score identifies drinkers and alcoholics with a view towards reliability. With the stigma attached to alcoholism, the respondent may be sensitive about questions related to drinking. Research has shown that the SMAST is minimally affected by denial tendencies (Selzer, 1975).

A series of twelve questions under the heading 'Your Family' were weighted and summed to

obtain a general family functioning score. It is supposed to reliably measure family functioning. The success or failure of interpersonal relationships within the immediate family might be thought of as another demographic characteristic of the individual. If there is some truth to the relationship between mental and physical health, covariates such as family functioning ought to be included in the study. In a similar vein, questions from the section 'Your Life in General' seek to measure social support outside of immediate family. An analogue to the family functioning score, the general well being score, was constructed.

Health outcomes may in part be determined by conditions in the workplace. I take a number of questions delving into workplace exposure to hazardous materials from this OHS section.

Lastly, physical activity has a direct impact on human health. Effects generally include the reduction of premature morbidity and an enhancement of emotional well being. The OHS asked about the type and frequency of physical activity that took place in the last month.

In conclusion, the OHS questionnaire obviously strives to be comprehensive. The sheer number of questions, totaling over one thousand, attests to the fact. With the availability of such data, this project has a good chance in uncovering a relationship as could reasonably be expected from an ecological study.

4.6 Survey Methodology

Survey methodologists implement sampling designs that meet given specifications under known constraints. In terms of a 95% confidence interval, the OHS design objective is to enable PHU proportions as low as 3% to be estimated within 50% of the estimate. For an estimated PHU proportion of 3%, the 95% confidence interval would be, in terms of percentages, $(3 - [1/2]3, 3 + [1/2]3) = (1.5, 4.5)$. Design constraints include budget and available sampling frames.

The chosen sampling frame, a frame of enumeration areas, comes from the 1986 Census. The enumeration area (EA) is the smallest area for which population counts can automatically

be retrieved. Each EA is situated in a PHU and classified as either urban or rural. Specifically, urban EAs represent the urban core and fringe of census metropolitan areas or census agglomerations.

A multistage stratified cluster sample is a good description of the OHS survey type. PHU and the urban/rural bifurcation stratify the population. The purpose of stratification is to group dwellings into homogeneous units with respect to measurable characteristics of interest. The estimates of population characteristics are for the most part more precise when using a stratified sample over a simple random sample. The primary sampling unit is the EA. In the first stage the survey takers sampled an average of 46 EAs within a PHU. They then constructed a list of dwellings for each of the selected EAs. The list became the sampling frame for the second stage of the sample. Clusters of dwellings, about fifteen from urban strata and twenty from rural strata were sampled at the second stage, resulting in the desired sample size of approximately 760 dwellings per PHU.

Reliable estimates of proportions greater than 3% had to be achieved at the PHU level. How was this criterion used to arrive at the sample size? Let $\hat{p}_\pi$ be an estimate of a proportion using weights determined by the survey design. Most designs for surveys conducted at a provincial or national level produce estimates with less precision than those that could be obtained by taking a simple random sample. The design effect (*deff*) of a proportion,, in this case estimated to be two (Ministry of Health, 1992a, p. 29), represents the factor by which the variance of an estimated proportion is inflated. Thus, letting Var $(\hat{p})$ be the variance of $\hat{p}$ under simple random sampling and ignoring the finite population correction factor,

$$\mathrm{Var}\,(\hat{\mathrm{p}}_\pi) \;\approx\; \mathit{deff}(\hat{\mathrm{p}}_\pi)\,\mathrm{Var}\,(\hat{\mathrm{p}}) \;=\; 2\,\mathrm{Var}\,(\hat{\mathrm{p}}) \;\approx\; 2\,\frac{\mathrm{p}(1-\mathrm{p})}{\mathrm{n}}.$$

The coefficient of variation is a scale free ratio of an estimate's precision to the expected value of the estimate. The OHS criterion for proportions greater than 3% was a coefficient of

variation less than 25%:

$$\text{C.V.}(\hat{p}_\pi) = \frac{\sqrt{\text{Var}(\hat{p}_\pi)}}{\text{E}(\hat{p}_\pi)} \approx \sqrt{\frac{2\,p(1-p)}{np}} < 0.25. \tag{1}$$

Since, by (1),

$$\sqrt{n} > 4\sqrt{\frac{2(1-p)}{p}} \qquad \text{or} \qquad n > 32\left(\frac{1-p}{p}\right)$$

and

$$\max_{p\in[0.03,1]}\left(\frac{1-p}{p}\right) = \frac{1-0.3}{0.3} \approx 32,$$

the approximate sample size needed to fulfill the reliability criterion is $n = 32 \cdot 32 = 1024$. For 46 PHUs the sample size translates to about 48,000 individuals.

Note that the sample size was further increased to account for expected nonresponse. The actual sample size resulted in 49,200 individuals responding out of 61,300 surveyed. The 49,200 individuals represent 35,500 households. To recapitulate, the large sample size exists to meet design specifications.

This section was included to give the reader a flavour of the methodological intricacies lurking behind the Ontario Health Survey data. The complex survey design induces nonequal probabilities of selection for the survey population; the weights associated with each survey respondent reflect selection probabilities adjusted for nonresponse and age-sex population totals at the PHU level. This should caution any analyst to consider carefully what types of inference are supportable by an analysis of the data.

4.7 Nonresponse Rates

Characterizing response rates for the OHS is not trivial. From the outset, the personal interview and self enumeration introduce at least two response rates: the OHS had a response rate of 88% for the former and 77% for the latter. Item response rates further fog the issue. I will document some of the difficulties in coming to terms with item nonresponse.

The topic of response rates is conventionally rephrased in terms of nonresponse rates. I will adhere to that scheme. Nonresponse can be divided into unit and item nonresponse.

Unit nonresponse occurs when the survey taker does not receive *any* information from the respondent. For the OHS, unit nonresponse happens if nobody in the household goes through with the personal interview. Unit nonresponse is handled operationally by modifying the probabilities of selection for the units selected in sample that do respond and, essentially, ignoring the nonrespondents.

Item nonresponse occurs when the survey taker procures only partial information about the respondent. Reasons for this type of nonresponse include respondent reluctance to answer sensitive questions and mistakes made during the transcription of data from the actual survey form to an electromagnetic file. For the OHS, item nonresponse is complicated by the fact that the survey is conducted using both personal interview and self enumeration. Thus, item nonresponse arises in the following scenarios:

Household Personal Interview	Family Member Self Enumeration
item nonresponse	unit nonresponse
item nonresponse	item nonresponse
item nonresponse	complete
complete	unit nonresponse
complete	item nonresponse

There are two common means of handling item nonresponse. The first is imputation. The technique makes up missing values. The two imputed OHS variables are age and sex. Missing values were imputed for the OHS by generating random values proportionally consistent with known PHU age-sex proportions. With imputed data the user cannot determine the rates of item nonresponse.

The second way of dealing with item nonresponse is to tell the user directly by allowing for a "not stated" category. This strategy allows for calculation of item nonresponse rates and the uncovering of nonresponse patterns. For example, the nonresponse for the eight questions on work exposure is given in Figure 5. Though the nonresponse rate hovers around 8% for each of the questions, for the most part the respondents either answer all of the questions or none.

Smoking nonresponse, shown in Figure 6, is an example of a more subtle pattern. The first

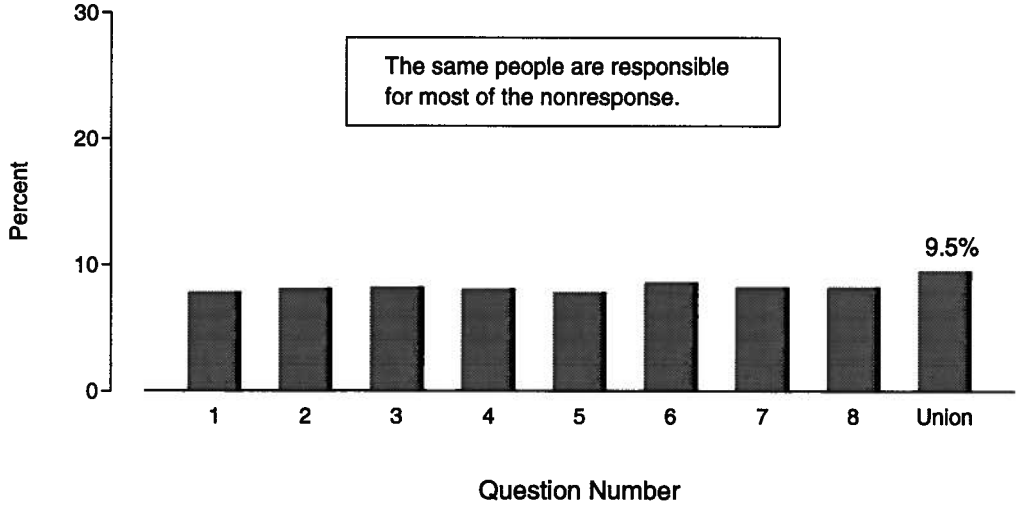

Figure 5: Nonresponse rates for the work exposure questions.

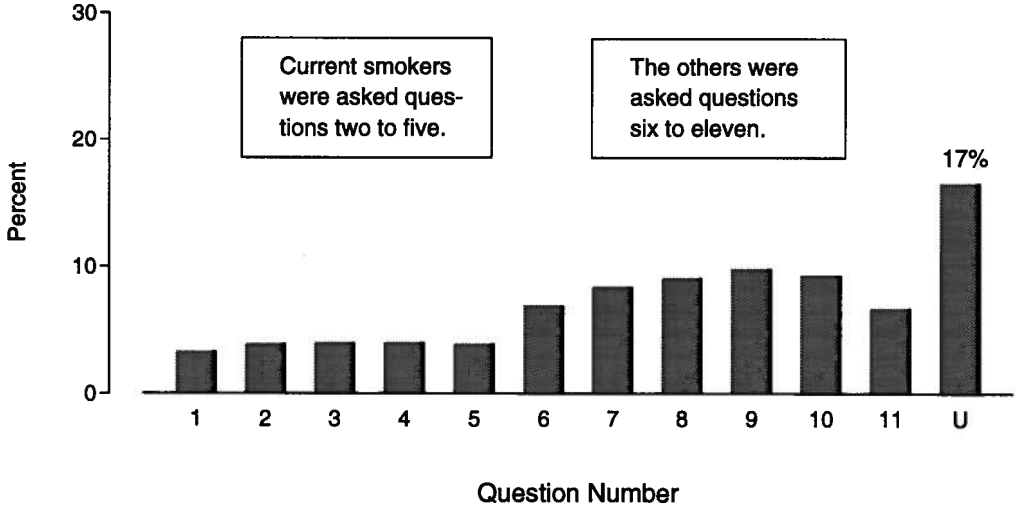

Figure 6: Nonresponse rates for the smoking questions.

Figure 7: Nonresponse rates for the well being score by age.

question determines if the respondent currently smokes. Current smokers are asked the next four questions while everybody else answers the six after that. The graph shows a higher item response rate for smokers but this merely reflects that a smaller proportion of the population can be classified as current smokers. Since the two groups are mutually exclusive the union for the response rate is approximately the sum of the current smoker and not current smoker nonresponse rate.

As a last example, consider the nonresponse pattern, shown in Figure 7, for a collection of questions concerning personal well being. Older respondents seem to be more sensitive to questions concerning their well being. The implication is that nonresponse is generally not random even though it is convenient to assume so for purposes of an analysis.

In conclusion, the Ontario Health Survey nonresponse is significant enough to warrant attention during analysis. What rates of nonresponse are there for the study variables? How will item nonresponse be dealt with for discrete and continuous variables? These types of questions will be dealt with as they arise.

4.8 Comment

The OHS comes with all the strengths and weaknesses of large scale survey data. The drawbacks include missing data and the survey weighting structure induced by complex survey methodology. Also, despite the wide subject coverage of the survey, the OHS provides no estimates of an individual's exposure to potentially harmful pollutants. We are left with the difficulty of estimating and linking pollution data from another source because of this absence. Admittedly, this criticism is somewhat unfair given the survey's objectives.

These drawbacks notwithstanding, I am delighted to have access to data backed by impressive resources, human and otherwise. One may think of the panels of experts who determined questionnaire content; those involved with the pretest; the survey methodologists; and the many who played a part in the field operations, from the training staff, interviewers and field supervisors to those completing the cycle with data capture and imputation. Untold hours went into the production of what is for me a starting point: the microdata file!

5 Initial Data Analysis

This section will give the reader an overview of the data used in the succeeding analysis. I draw data from two sources: the 1991 Ontario Health Survey (OHS) and six years of atmospheric environmental monitoring.

The OHS target population (minus immigrants who have lived in Ontario for less than ten years) comprise the study population. We exclude recent immigrants because our outdoor air pollution estimates would not adequately represent their true exposure history.

I am forced to consider individuals as the unit of analysis because the OHS public datafile, restricted for reasons of confidentiality, does not identify their household. A conflict immediately arises since the analysis should be in synchronicity with the survey design, meaning that households rather than individuals should be the unit of analysis. One implication of employing standard estimation techniques is that standard errors will be underestimated if no adjustments are made.

In its complete form the OHS data is unmanageable. There are over 1000 variables of which many are of no use to this study. The first task is to cull the data. I give a qualitative and graphical description of the reduced set of variables along with associated nonresponse rates.

The pollution data I start with derive from an involved inferential process. The original data comes from thirty seven atmosphere monitoring stations. Each station potentially measures up to four pollutants, namely nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2) and sulfates (SO_4). The station data is interpolated for each public health unit (PHU) by Weimin Sun to whom I am indebted.

I convert the monthly averages, given over the six year period 1983-89, into single summer and winter averages. Thus, each PHU has a six year average which I assume adequately represents personal lifetime pollution exposure. For a flow diagram of the way the study data is derived see Table 5.

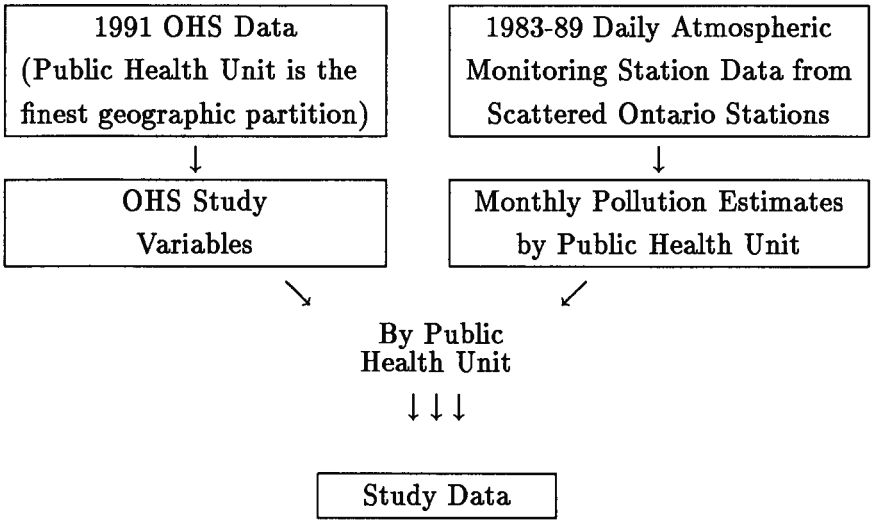

Table 2: Flowchart of the derivation of the study data.

To close, I graph each of the covariates against the two response variables. This will provide intuition about at least the one term models we fit later.

5.1 OHS Covariates

The OHS data contains information on innumerable aspects of the population of Ontario. Since these data derive from measuring over 1000 variables, a subset that will best relate pollution to respiratory illness must be selected. In this section I will describe what covariates were selected, report their associated nonresponse rates and illustrate their marginal distributions graphically.

This study focuses on asthma and emphysema as the response variables. The two questions from the interviewer portion of the questionnaire were:

"Do you have asthma?"; and

"Do you have emphysema or chronic bronchitis or persistent cough?".

The questions assume an ongoing chronic condition by the way in which they are asked. A small fraction of questionnaire respondents (0.9%) did not respond to the two questions specifically. This fraction will be ignored from here on. Figure 8 illustrates the prevalence of asthma and

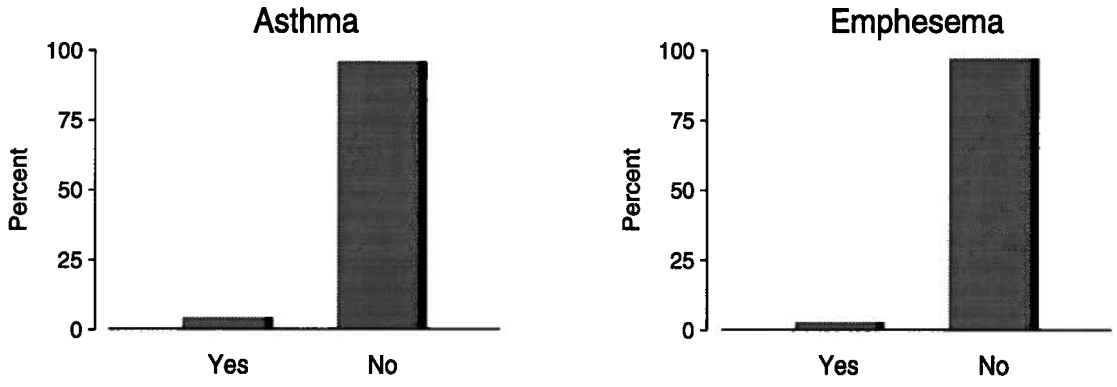

Figure 8: Prevalence of the two response variables.

emphysema. The low rates observed point to the need for a large sample: such studies could be straight-jacketed by the lack of statistical power resulting from small samples.

To select the predictive covariates, I tried to obtain a satisfactory coverage of the following set of individual descriptors: demographic, lifestyle, health, socioeconomic, and pollution exposure measures. The categories are rather arbitrary, though I chose them to make the presentation of results more comprehensible. Table 3 exhibits the set of variables I selected in each group. The grouping adhered to in the table will generally not make any difference to the outcome of reported results; the arcsine analysis is the only exception.

Most of the study variables have been massaged in a variety of ways. The producers of the OHS datafile made the preliminary alterations. To make the file easier to use, they combined subsets of the original questions to get *derived* variables. Household type, for example, classifying the respondent according to a description of the relationship between the family members of the household, was derived from the Household Record Form for each family in each household.

Household income is another example. The variable classifies the respondent into one of three household income groups: low income; not low income but less than $50,000; and income $50,000 or more. A question that arises is how exactly low income is defined. The cost of living, household size and household income are determinants that should be taken

Grouping	Covariate	Type
Demographic	Rural or Urban Stratum	Nominal
	Sex	Nominal
	Age	Continuous
	First Generation Immigrant	Nominal
Socioeconomic	Family Type	Nominal
	Family Functioning Score	Continuous
	Blue Collar Work	Nominal
	Work Exposure	Ordinal
	Post-Secondary Education	Ordinal
	Income	Ordinal
Lifestyle	Smoker Type	Ordinal
	Current Smoker Type	Nominal
	Duration Smoked	Continuous
	Number of Cigarettes Smoked	Integer
	Number of Current Household Smokers	Integer
	Alcohol Problem	Ordinal
Health	Body Mass Index	Continuous
	Energy Expenditure	Continuous
	Well Being Score	Ordinal
	Allergy	Ordinal

Table 3: List of study covariables.

into account. The rule adopted by the OHS is based on poverty lines and low income cut-offs developed by the National Council on Welfare and Statistics Canada. Place of residence (urban or rural), income and household size determine low income classification (Ministry of Health, 1990, p. 11).

I introduced the second set of modifications which are typified by the example in Table 4. In effect I used existing OHS variables to derive a new variable. I chose this route to reduce the number of explanatory covariates under consideration. Clearly the choice was arbitrary to some degree.

The last modification of the original OHS data is related to item nonresponse. I will delay my exposition of nonresponse until after I have described the explanatory covariates in more detail.

Questions:	In your job or business have you, in the past twelve months, worked with	Responses:
1.	dust from wood, grain, haw or straw?	
2.	dust from silica, granite or rock dust?	Never
3.	glass fiber dust or asbestos?	Occasionally
4.	dust or fumes from lead, cadmium, nickel, chromium or mercury?	Often
5.	fumes from solvents, paints or gasoline?	Always
6.	resins or isocynates?	Don't Know
7.	pesticides?	Not Applicable
8.	coal tar or pitch?	Not Stated

$$\Downarrow \qquad \Downarrow \qquad \Downarrow$$

$$\text{Work Exposure} = \begin{cases} \text{Yes} & \text{if 'Often' or 'Always' at least once,} \\ \text{No} & \text{otherwise.} \end{cases}$$

Table 4: How the OHS work exposure questions form a covariate for this study.

The demographic covariates describe certain unalterable features of the respondent. I chose stratum, sex, age and an immigrant indicator as the demographic covariates. Their marginal distributions are shown in Figure 9. As with the other estimates in this chapter I used survey weights, as prescribed by OHS documentation (Ministry of Health, 1990(c), p. 3.), to produce the estimates.

The four demographic covariates look reasonable. The *Canada Year Book* (Statistics Canada, 1991, p. 73) tells us that 83% of the Ontario population resides in an urban setting. The OHS weighted estimate is 86%, as expected. The division between the sexes is about fifty-fifty and the age distribution shows a bulge for the baby boomers. The immigrant indicator shows what portion of the study population are immigrants who arrived before 1980.

Age may be an important explanatory variable. The probability of death increases with age, ranging from 3% for Canadians between the ages of one to 24 to 71% from 65 years and up (Future Health, p. 102). Any disease related to chronic exposure over a long period of time might be expected to be related to age. For ailments of the respiratory tract, older people

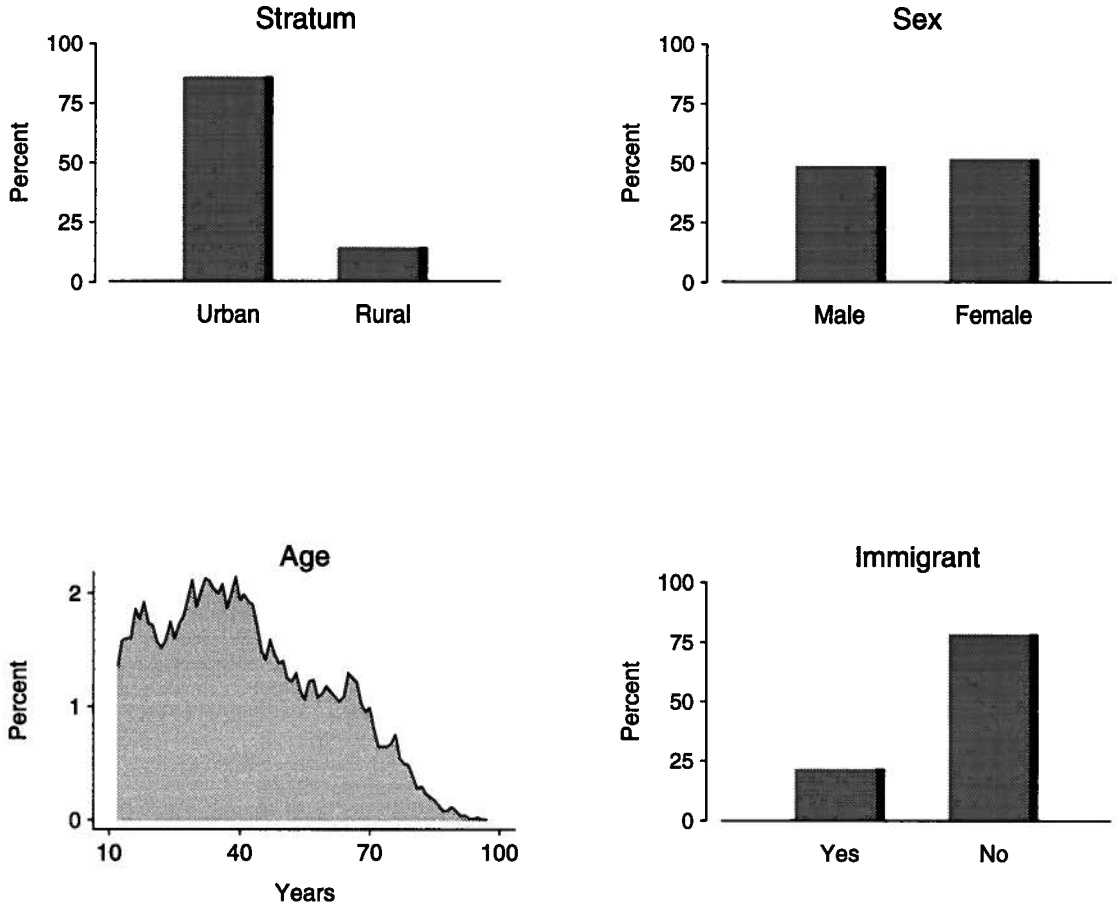

Figure 9: The demographic covariates.

appear to be more vulnerable to inhaled particles and gases (Brain, 1989).

Some studies consider ethnicity as a demographic characteristic. When ethnicity proves to be a good predictor, however, it likely reflects class membership, as in the case of aboriginals. As a group, chronic cough or emphysema affects them at almost double the Ontario average (4.5 versus 2.4%). This phenomenon is explained by the well documented condition of poverty in which many natives live. The OHS data reinforce this view with an observed moderately high negative correlation between chronic cough and standard socioeconomic variables such as education and income. The higher prevalence of illness therefore reflects structural inequalities within society rather than race (Steinberg, 1984). Since ethnicity in Canadian society, exempting aboriginals, does not generally imply class membership, this study relies on the

socioeconomic indicators to capture those inequalities.

Socioeconomic variables, shown in figure 10, are important indicators of longevity (see references in Evans, Tosteson and Kinney, 1984). Presumably they are also good predictors of respiratory health. The socioeconomic covariates represent information on the family unit, work, education and income. All of these variables are categorical, including the seemingly continuous family functioning score. The score, however, is a weighted score of responses from a series of twelve questions from the self completed questionnaire. The actual cutoff is supposed to represent the best division distinguishing families seeking clinical help from those in the general population. The income categories definition depends on poverty lines and low income cutoffs developed by the National Council of Welfare and Statistics Canada. The formula adjusts for household size, area of residence (urban/rural) and household income (Ontario Ministry of Health, 1992a, p. 10, 21-22).

Lifestyle covariates, shown in Figure 11, may turn out to be the most important set of explanatory variables due to the impact of smoking on lung function. The best of them is probably duration smoked as it is continuous and more reliable that the other continuous covariate, the number of cigarettes smoked daily. The measurement of the number of cigarettes smoked daily illustrates the tendency of people to 'think in fives' when asked for a simple answer to a complex habit. Second hand smoke exposure was captured in the personal interview when the respondent was asked if *anybody* in the household smoked. Finally, a drinking problem index, developed from a shortened form of the Standard Michigan Alcohol Screening Test (SMAST) is included as one of the lifestyle indicators.

The general health of a person may have an effect on specific pathologies such as asthma or emphysema. Figure 12 illustrates the distributions for the health covariates. Body mass index (kg/m^2) and exercise expenditure ($kcal/kg/day$) are continuous and reflect individuals' participation in physical exercise. Well being, for reasons similar to the family functioning score, is ordinal. The family functioning variable divides families into functional and dysfunc-

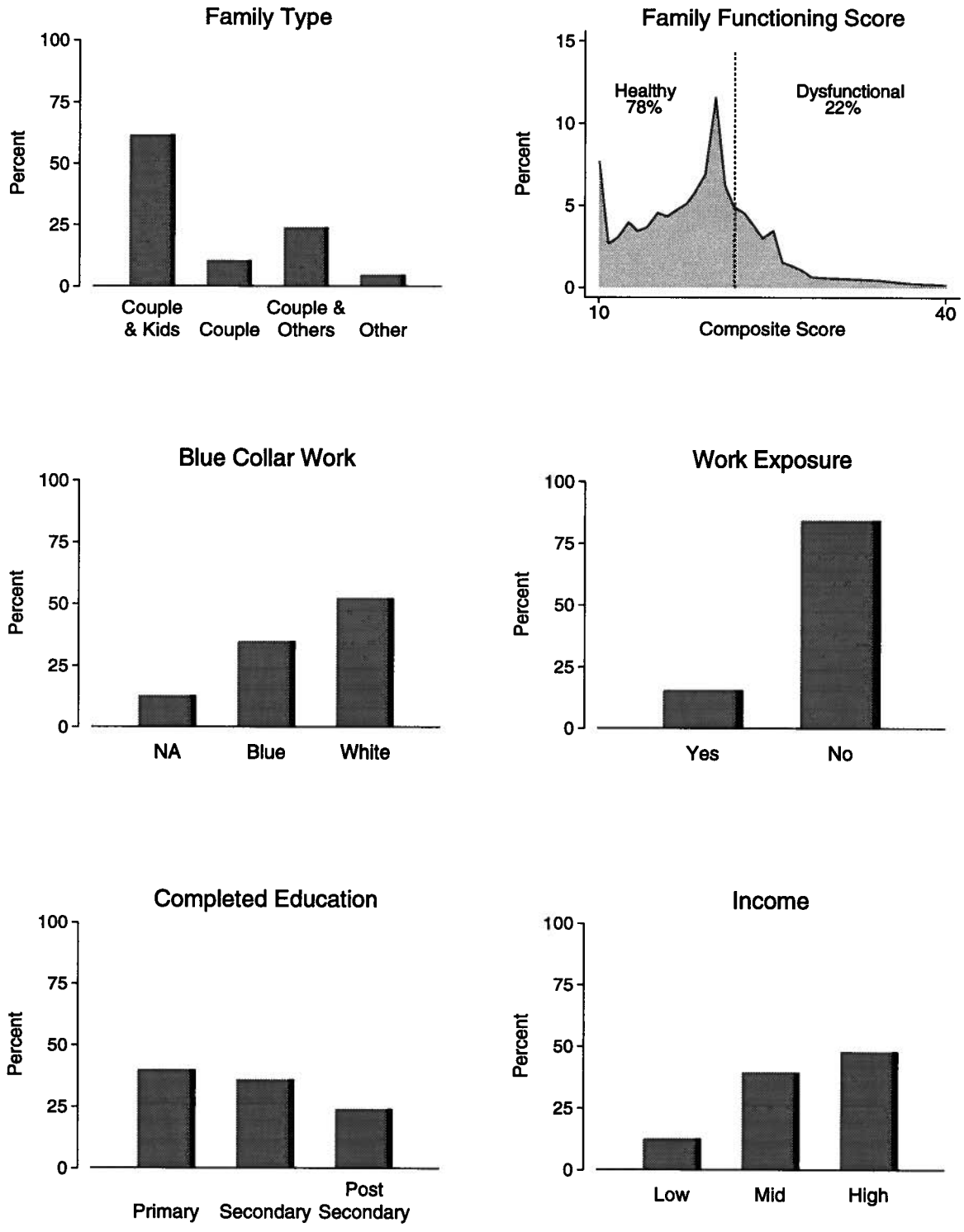

Figure 10: The socioeconomic covariates.

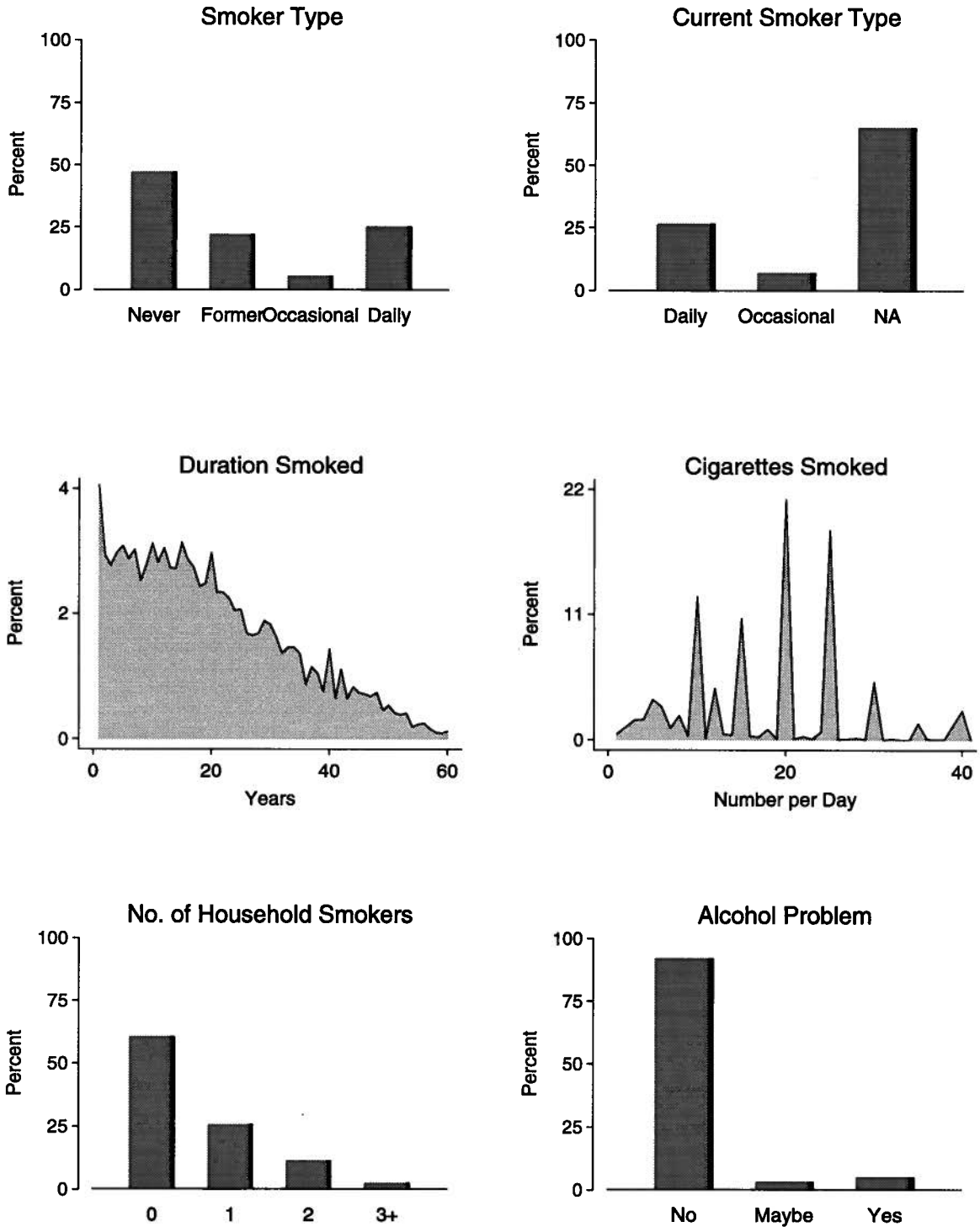

Figure 11: The lifestyle covariates.

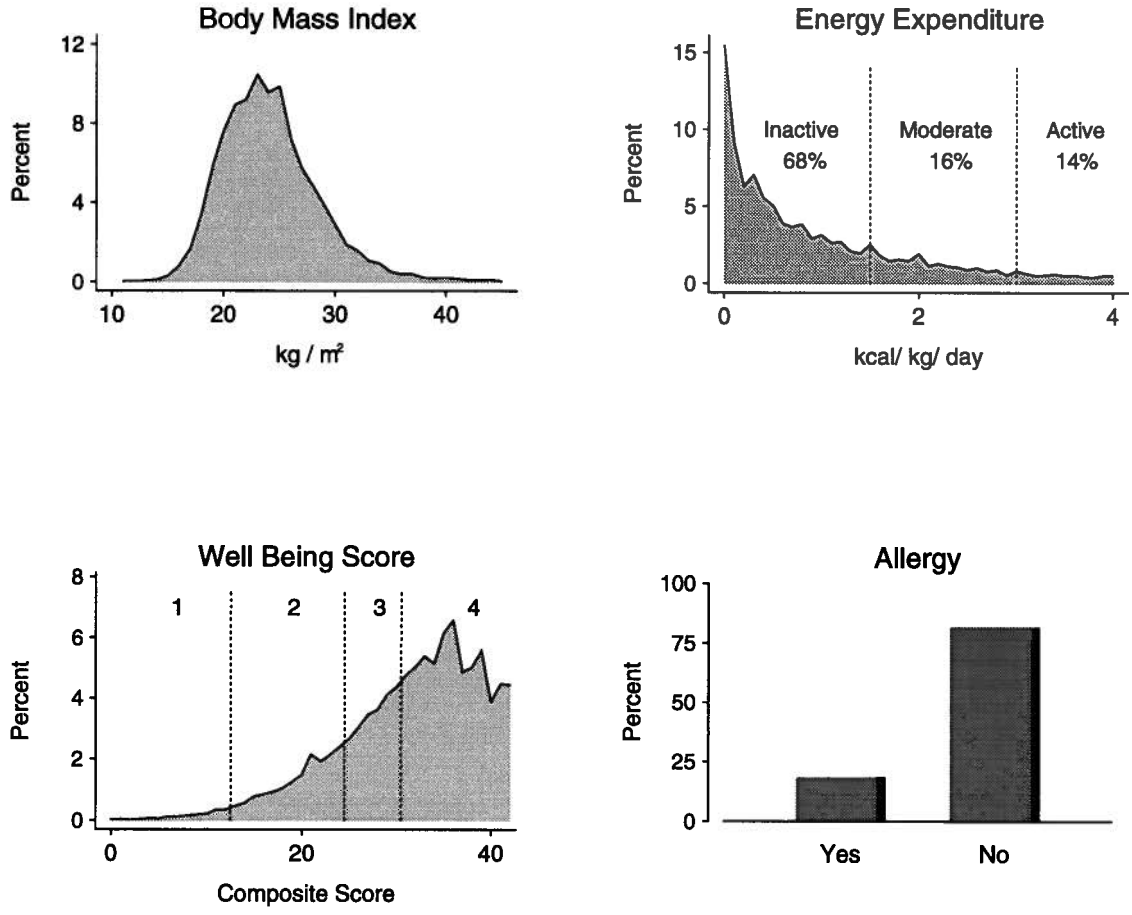

Figure 12: The health covariates.

tional.

Most of the OHS covariates display some degree of nonresponse though for the graphs given so far, the nonresponding portion of the sample was ignored. I will currently take some space to characterize the nonresponse for the study data.

Nonresponse can be divided into unit and item nonresponse. In the context of the OHS, unit nonresponse was either at the *household* level, where no personal interview took place, or at the *household member* level, where an individual's self completed questionnaire was not obtained. Item nonresponse occurred when a respondent provides partial information. Here, item nonresponse manifests itself as a respondent who partially omitted information on the self completed part of the survey.

The OHS had a unit response rate of 88% for the personal interview and 77% for the self completed questionnaire. For the rest of the study I will consider only the 77% of respondents responding to the self completed questionnaire.

Figure 22 in the Appendix summarizes item nonresponse. The range is rather striking. Hovering around 15% are well being, energy expenditure and household income. For the next set of variables, from the number of cigarettes smoked to smoker type, item nonresponse stands at about 10%. Blue collar work, allergy, education and immigration show negligible nonresponse.

There are at least three plausible explanations for high nonresponse. First, the respondents' sensitivity to the question has an effect on rates. The question on household income is a sterling example. Many Canadians may well feel uneasy about providing the information. For some the anxiety is culturally related; for others revealing income may be embarrassing; yet others may feel the government could use the survey as a device to nab tax evaders. Well being is another example: observed nonresponse increases as age increases.

Second, the length of the questionnaire will have an effect on both unit and item nonresponse. In the OHS pilot study, a short version of the survey resulted in higher overall response levels. There is reason to believe that item nonresponse is also affected by questionnaire length. The fact that some household members skipped whole parts of the survey is suggestive.

Third, some variables are composite indicators and nonresponse can then result from a missing answer for only one question. The well being score is a weighted total of one positive and one negative statement covering seven categories: energy, control of emotions, state of morale, interest in life, perceived stress, perceived health status and satisfaction about relationships. If any one of fourteen questions went unanswered the well being score was coded as "not stated". This phenomenon may explain the two other composite variables, family score and energy expenditure.

Some of the variables seem to have achieved 100% response rates. In the case of the

Exposure Assessment	Effects Assessment
Level	Hazard Identification
Distribution	Type of Effect
Number of People	Dose Response
Target Dose	Risk Characterization

Table 5: Types of pollution assessment studies.

geographic indicators Public Health Unit (PHU) and stratum, data exists for all respondents because the variables were used in survey stratification. In other cases the response rate is artificial. An imputation method can, for example, fill in data where data is missing. Age and sex were imputed using a random assignment mechanism based on census information on a stratum's population breakdown into age and sex categories.

From this look at item nonresponse we can agree that some of the variables are more reliable than others. Looking at the set of study variables together, only 55% of the respondents have given complete information. At the modelling stage provisions must be made to deal with the item nonresponse.

In summary the 1991 Ontario Health Survey is the source of the two response variables and twenty explanatory covariates. The majority of covariates are categorical rather than continuous, reflecting the difficulties of obtaining good continuous measurements from a large scale survey. The set of explanatory variables cover a range large enough to build decent models. The one ominous omission is the pollution data to which I now turn.

5.2 Pollution Covariates

The measurement of internal dose of a pollutant is very important in pollution effects studies. In most cases, however, the internal dose is unknown. The theme of this section is how the available exposure data is related to an individual's internal dose.

Consider the two broad classes of pollutant assessments shown in Table 5. The most commonly available data measures environmental releases or concentrations in specific media;

exposure measurements such as the number of people exposed and their absorbed dose are relatively rare. Exposure data is therefore generally estimated by making assumptions that will allow pollution concentration data to be linked with individuals falling in a certain geographical area. For most pollution assessment studies the data is clearly imperfect. This study falls under the effects assessment heading as a risk characterization study. How good the assumptions needed are is uncertain but at least one study, where NO_2 exposure measured by personal monitoring devices was compared to station measurement, suggests station measurements are adequate for airborne pollutants (Hackney et al., 1992).

The contact between a person and a pollutant in an environmental medium is termed exposure. Exposure is completely described by the *route* by which the pollutant enters the body; the *concentration* of the incoming pollutant; the *duration* of the exposure; and the *frequency* of exposure. Most pollutant data measure pollutant concentration for one medium covering some geographic area (Sexton et al, 1992).

I make the following assumptions about the pollution data:

- ▷ The weather station sites have well calibrated measuring instruments. That is, at a given concentration of SO_4, say, all stations would give a reading closely corresponding with other stations.
- ▷ A six year average of site pollution is a good indicator of longer term averages.
- ▷ Average pollution levels do not fluctuate radically from decade to decade.
- ▷ Spatially, ambient aerosol pollution is homogeneously distributed within Public Health Units (PHU).
- ▷ Ambient aerosol pollution levels are proportional to an individual's absorbed internal dose.
- ▷ Alternative sources and environments for the pollutant under study are negligible.

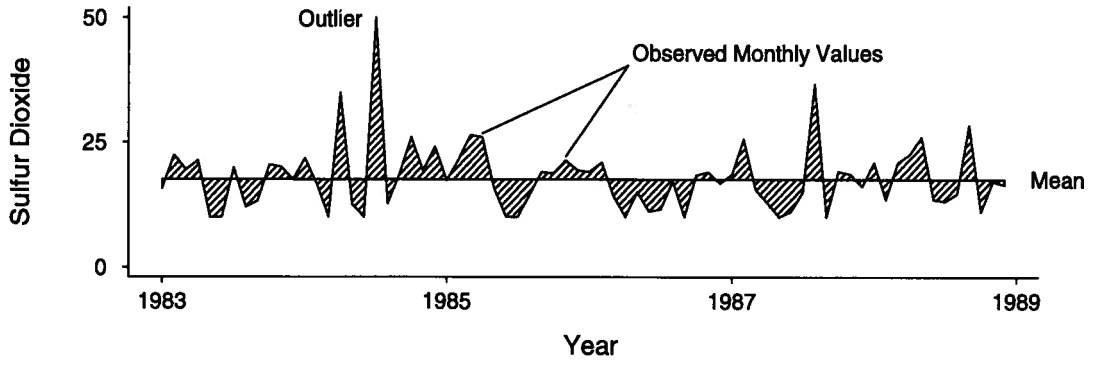

Figure 13: Typical graph of a station's measurement of one pollutant.

▷ Individuals within a PHU, the approximate equivalent of a Census Division, are spatially stable; migration from one PHU to the next is minimal.

These assumptions make the available data appropriate for our analysis. The data come from 37 atmosphere monitoring stations unevenly scattered across Ontario. Not all stations measure all pollutants of interest.

The station data is used to predict pollution levels for the PHU centroids for which there are no stations. Brown, Nhu and Zidek (1993) have developed a methodology for spatially interpolated predictive distributions. The modification and implementation of their method, described in Duddek et al. (1994), produced the predicted monthly averages used in my analysis. In this section I will heuristically explain the steps taken to get the pollution estimates. The first step begins with the original measurements from which the predicted means are eventually estimated.

Nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2) and sulfate (SO_4) are the four pollutants considered in this study. The daily measurements, taken over the six year period, 1983-89, have been converted into monthly averages. Figure 13 is an example of a measurement of one pollutant at one station and indicates how the graphs are to be understood. The more complete sets of graphs are given in the appendix by Figures 23 to 26.

Station measurements of NO_2 are given in Figure 23. The plots of twelve of the thirteen stations are ordered by ascending station averages. By looking at the first and last stations one gets a sense of the mean range, in this case from 25 to 45 $\mu g/m^3$ NO_2. Low outliers appear in graphs eight and ten. Their existence is partly explained by many missing daily measurements. If at least one daily measurement is available for the month then the monthly mean is calculated. Nothing is done about measurement or transcription error since no information on data quality is available. If no measurements exist for the month a monthly mean is imputed. On the whole, the effect of outliers on the six year mean is minimal; the observed difference between the simple average and winsorized mean eliminating one observation on each extreme is less than five percent.

Ozone measurements age given in Figure 24. The ozone data is better than the NO_2 data in two respects. First, there are almost twice as many stations measuring ozone as NO_2. Other conditions being equal we have more information available for ozone. Second, in contrast with NO_2, a temporal pattern with peaks in summer and lows in winter is evident. This allows me to more easily identify poor stations by observing which deviate from the trend. The data from the station in the fourth row and the third column, for example, looks somewhat suspect.

Sulfur dioxide and sulfate measurements are next. Twice as many stations measure SO_2 as SO_4. No seasonal pattern appears in either. SO_2 has the highest between station variability among pollutants.

Within each PHU we choose a centroid for which a monthly estimate will be produced. The estimate is supposed to be a good proxy for the internal dose experienced by people living in that PHU. Our interest lies in estimates of six year averages, broken down by summer and winter. By dividing the estimates into two types we had hoped to capture subtle spatial differences that may exist between seasons. The final pollutant estimates are shown in Table 6.

We make two checks on the validity of the estimated data. First, for each pollutant we

Public Health Unit	Summer				Winter			
	O_3	NO_2	SO_2	SO_4	O_3	NO_2	SO_2	SO_4
Eastern Ontario	45.9	26.6	22.5	4.04	30.3	31.3	22.8	4.30
Ottawa-Carleton	46.0	25.5	22.8	4.33	30.0	31.5	22.1	4.42
Leeds, Grenville and Lanark	46.1	25.5	19.7	4.75	29.5	30.8	22.1	4.55
Kingston, Frontenac, Lennox	46.3	25.6	18.0	5.24	28.5	30.5	21.2	4.51
Hastings and Prince Edward	45.9	25.6	17.6	5.26	28.4	30.9	20.7	4.54
Haliburton, Kawartha, Pine Ridge	46.3	28.2	16.9	4.88	29.1	33.1	18.8	4.78
Peterborough	45.8	26.0	17.0	5.23	28.3	31.5	19.7	4.68
Durham	48.1	30.7	16.6	5.89	27.9	32.8	19.2	4.92
York	48.0	33.3	15.4	4.97	29.0	34.7	18.3	4.85
Toronto	49.8	37.4	18.6	5.99	27.6	34.6	20.2	4.91
Peel	48.7	36.1	16.0	4.89	28.7	35.7	18.7	4.73
Wellington, Dufferin and Guelph	49.0	37.8	16.2	4.37	29.6	37.4	19.0	4.66
Halton	50.4	38.4	17.8	5.48	28.0	35.5	20.0	4.76
Hamilton-Wentworth	52.5	36.1	18.4	6.17	28.8	33.9	20.3	5.10
Niagara	56.4	24.1	18.0	8.73	30.9	27.7	20.3	6.10
Haldimand-Norfolk	57.5	25.5	18.3	8.75	31.1	28.7	20.1	6.25
Brant	52.6	35.7	18.0	6.01	29.0	34.2	20.2	5.05
Waterloo	49.4	39.7	16.5	4.67	28.4	37.7	19.8	4.47
Perth	49.5	40.1	18.7	4.33	29.0	39.0	21.5	4.40
Oxford	52.7	36.1	18.3	5.76	29.1	35.3	20.5	4.91
Elgin-St.Thomas	56.5	30.1	20.4	7.60	30.3	32.7	20.9	5.57
Kent-Chatham	57.8	29.5	22.1	8.46	30.2	35.2	22.5	5.48
Windsor-Essex	58.9	28.4	22.9	9.23	30.1	36.8	24.1	5.51
Sarnia-Lambton	55.3	37.0	29.1	5.89	30.0	39.0	25.3	5.11
Middlesex-London	53.8	36.6	21.9	5.73	29.4	37.1	22.3	4.86
Huron	49.4	40.5	20.3	4.10	29.8	39.6	22.8	4.52
Bruce-Grey-Owen Sound	47.5	37.1	16.1	3.63	31.7	37.7	19.0	4.66
Simcoe	47.2	33.1	15.0	3.93	31.0	36.1	17.8	4.54
Muskoka-Parry Sound	44.7	31.8	17.7	3.27	28.5	36.0	23.2	2.90
Renfrew	44.9	28.3	19.5	3.79	29.8	34.4	20.3	3.57
North Bay	42.8	31.4	20.0	2.80	27.9	34.2	21.9	2.56
Sudbury	38.6	33.3	33.3	1.96	26.4	35.7	30.1	2.15
Timiskaming	43.5	32.3	25.1	2.63	28.0	34.0	23.6	2.47
Porcupine	47.5	32.2	21.9	3.19	28.9	33.5	22.5	2.61
Algoma	39.0	34.2	30.3	2.00	26.6	36.3	28.6	2.16
Thunder Bay	43.7	33.9	24.7	2.49	27.0	35.3	25.3	2.11
Northwestern	49.0	32.2	20.8	3.39	29.7	33.0	21.5	2.78

Table 6: Spatially interpolated ambient air pollution six-year averages ($\mu g/m^3$).

produce boxplots for estimated and actual values (Figure 29 in the Appendix). The range of observed values may be larger than the estimated values, in particular for nitrogen dioxide and sulfur dioxide, because some stations had high measurement variability. The spatial interpolation method corrects for that variability by emphasizing stable stations over highly variable ones.

The second check comes in the form of a display of estimates by PHU in Ontario. Figures 30 to 33 exhibit the distribution of pollution estimates for both summer and winter. Dark areas indicate higher pollution levels than lighter ones. The gradual nature of regional differences and higher estimates for the more heavily industrialized Great Lake PHUs give us confidence in the method.

Now that we have estimates for both summer and winter we must address whether it is worthwhile to keep two variables for one pollutant. If summer and winter pollution estimates are highly correlated then a combined pollution estimate should suffice. The strongest relationships between all summer and winter combinations for all four pollution variables is shown in Figure 27.

The linearity of the top three graphs stands out at once. Since the three pollutants, nitrogen dioxide, sulfur dioxide and sulfates, have similar ranges, data averaged over the year will probably do just as well as either of the two measures. Ozone is rather quirky with summer values impressively larger in the four month summer than in the eight month winter. Since people are generally outside more in the summer, the summer ozone value will be the ozone representative value in the analysis.

The last five graphs show the strongest relationships between pollutants of differing types. Significantly, ozone shows up in each of the plots. Summer ozone is positively correlated with summer sulfur dioxide and, strangely enough, winter sulfates. On the bottom row, winter ozone shows a hint of a relationship with summer sulfur dioxide, winter sulfur dioxide and winter sulfates. From the limited data, nothing interesting shows up for nitrogen dioxide.

That concludes our description of the process from original station measurements to six year PHU estimates. We must now link the pollution data to the Ontario Health Survey data and proceed to the analysis.

5.3 Asthma versus Covariates

Once the pollution data is merged with the OHS data we are poised to start with the preliminary stages of the analysis. We can make a first assessment of the relationship between asthma and each of the explanatory covariates by looking at their plots. (Figures 14 and 15).

The bars in the plots exhibit two characteristics. The first is the whitespace in the middle of the bar, the weighted estimate of the proportion. The second is the length of the bar, primarily indicating the number of observations used to calculate the proportion. The bar length, however, is not an explicit 95% confidence interval. A nominal 95% confidence interval would generally be the weighted proportion plus or minus two times the binomial variance. With survey data we have the design effect (*deff*) mentioned in the description of OHS methodology. Since the OHS documentation suggests that the design effect for an estimate of the mean is around two I have constructed the following intervals:

$$\hat{p}_\pi \pm deff(\hat{p}_\pi)\ z_{0.95}\sqrt{\frac{\hat{p}_\pi(1-\hat{p}_\pi)}{\hat{n}_\pi}} \approx \hat{p}_\pi \pm 2\cdot 2\sqrt{\frac{\hat{p}_\pi(1-\hat{p}_\pi)}{\hat{n}_\pi}}$$

Of course, by displaying so many plots we are implicitly making multiple comparisons. I have not attempted to construct intervals that maintain an overall 95% confidence level but rather have plotted them in the spirit of an exploratory analysis.

Of the demographic variables, age is the most interesting. Asthma has a higher prevalence for the youngest members of the population. This suggests that some proportion of asthmatics are relabeled later on in life.

In the socioeconomic sphere household type, education and income show mild correlation with asthma. The household type 'D' represents single parent households, a category known to have a disproportionate number of poor, single women. The fact that household type,

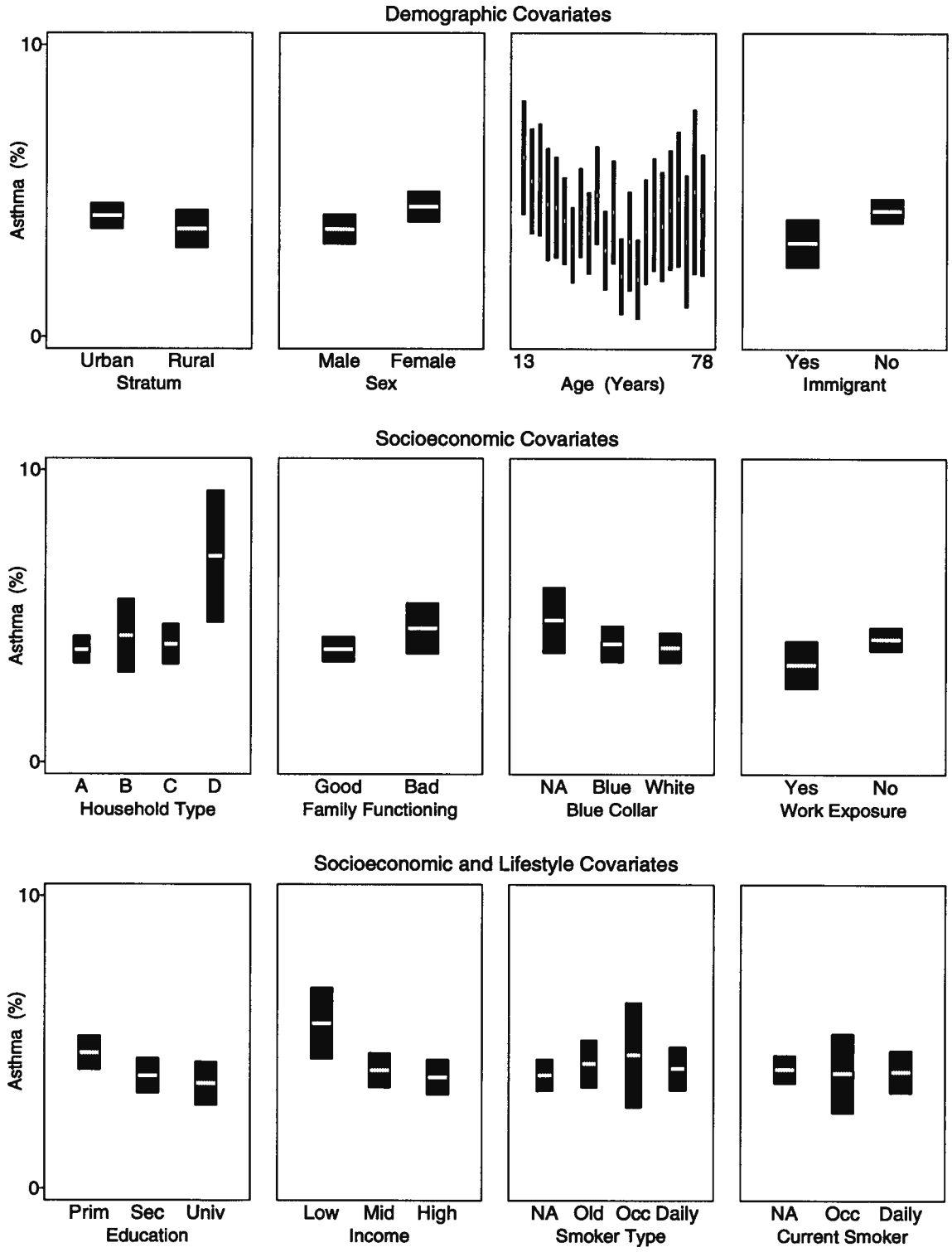

Figure 14: Asthma prevalence by covariates (I).

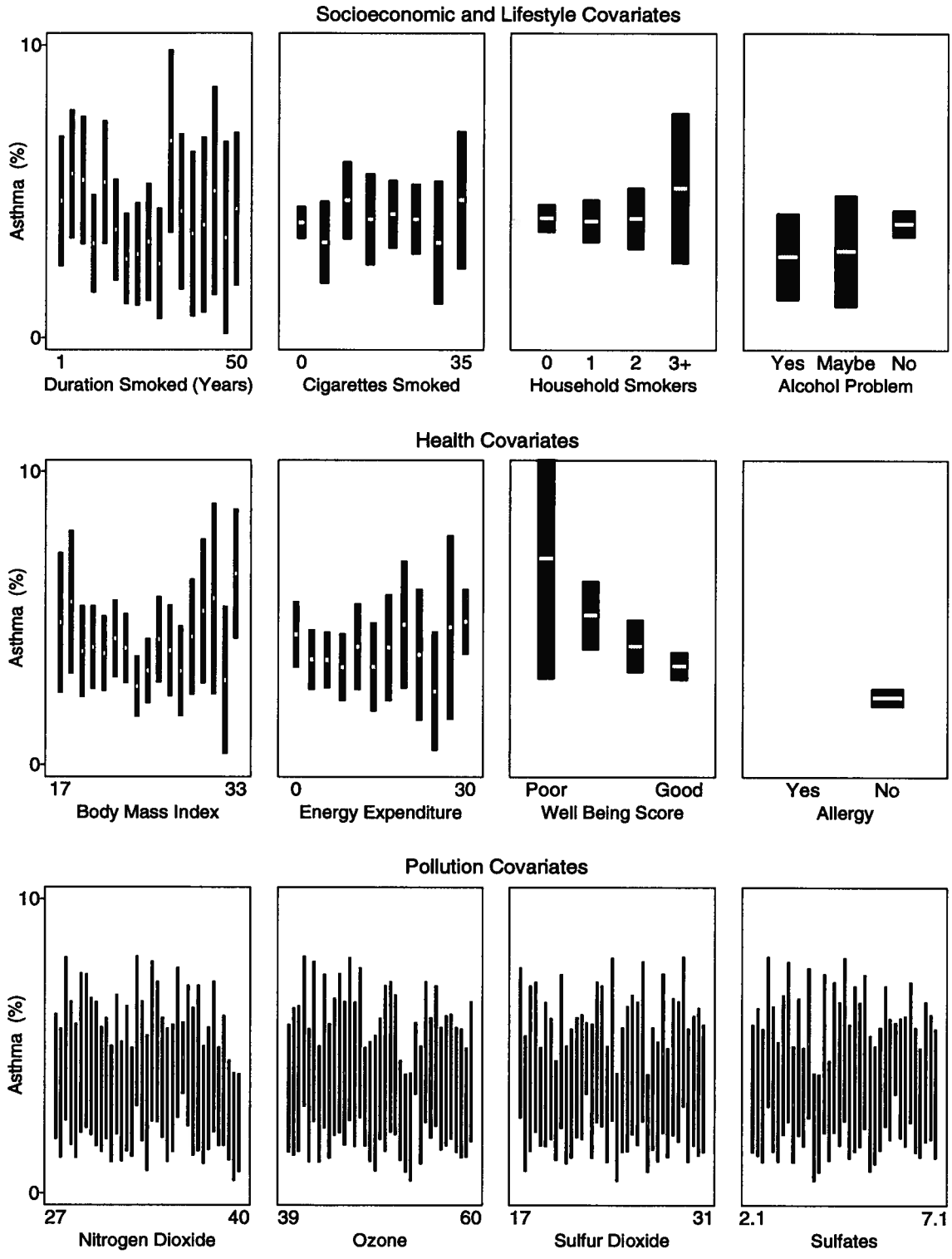

Figure 15: Asthma prevalence by covariates (II).

education and income concur strengthens the argument that socioeconomic factors are related to health outcomes.

Work exposure merits some attention: observe the lower prevalence of asthma for the population exposed to dust and fumes in an occupational setting. The most likely cause is that asthmatics probably try harder to avoid work that would exacerbate their condition. This is a perfect example of the caution that must be heeded in making inferences with observational data.

The smoking variables display no interesting relationships but two of the health variables do, ie. the well being score somewhat and the allergy indicator strongly. The presence of well being and allergy poses a problem to the analyst wishing to use them to explain asthma or emphysema. Are they reasonable covariates if they in fact are another measure of respiratory ailments? On the other hand they could be thought about as a way of significantly reducing the variation in the model and thereby increasing the power of hypothesis tests. The position I will take is to avoid bringing them into the model due to the problem of interpretation. It would hardly be right to say that a feeling of being ill brings on asthma.

Finally, the lack of pollution trends is revealing. Each estimated proportion is based on an average of over one thousand sampled individuals living within a PHU. The graphs raise questions about how any meaningful relationship between asthma and pollution could be discovered.

In summary allergy shows the strongest relationship though its use as an explanatory covariate is questionable. Age, education and income show slight downward trends. In all there are no compelling signs to indicate that we can fit a good model for asthma as the response variable.

5.4 Emphysema versus Covariates

In the same spirit Figures 16 and 17 show the prevalence of one or more of bronchitis, chronic cough or emphysema against each of the covariates. It is immediately obvious that the esti-

mated prevalence rates exhibit greater differences here than for asthma.

As age increases so does the prevalence of emphysema. Perhaps this finding is not too surprising since the lungs lose some of their elasticity as time passes. None of the other demographic variables are helpful.

Of the socioeconomic variables education and income reiterate the oft observed relation between socioeconomic status and health. The poorer one is the more likely one will be afflicted with medical conditions less prevalent in the upper socioeconomic order.

The smoking variables are strong explanatory covariates. Smoker type and current smoker status show us what we would expect given current knowledge: smoking is detrimental to the functioning of the respiratory system. Duration smoked and, to a lesser extent, the number of cigarettes smoked show distinct upward trends. The effect seems most pronounced for the long time smokers, ie. the upper third of smokers. Functionally, a quadratic equation looks like it would fit best. Second hand smoke, partially measured through the number of household smokers, shows a slight upward trend as does high alcohol consumption. Undoubtably any modeling of emphysema must incorporate at least one smoking covariate.

The measures of health status display some of the correlation we expect to see. Energy expenditure, well being score and allergy have gentle downward slopes. Correlation between the health covariates and smoking status could wipe out any of the effect we might attribute to the variables (also see the reservations stated in the previous section).

The pollution covariates are again negative. The bars look as if they are randomly strewn over the range of estimated pollution values. The only hope in relating any of the pollution covariates to emphysema will be to model out as much of the variation inherent in the estimates and condition on other covariates. That is the focus of the next two chapters.

Unlike asthma, we anticipate that the modeling of emphysema will at least result in models with descriptive power; age, duration smoked and income are related marginally to emphysema. Whether pollution will show significance, however, remains to be seen.

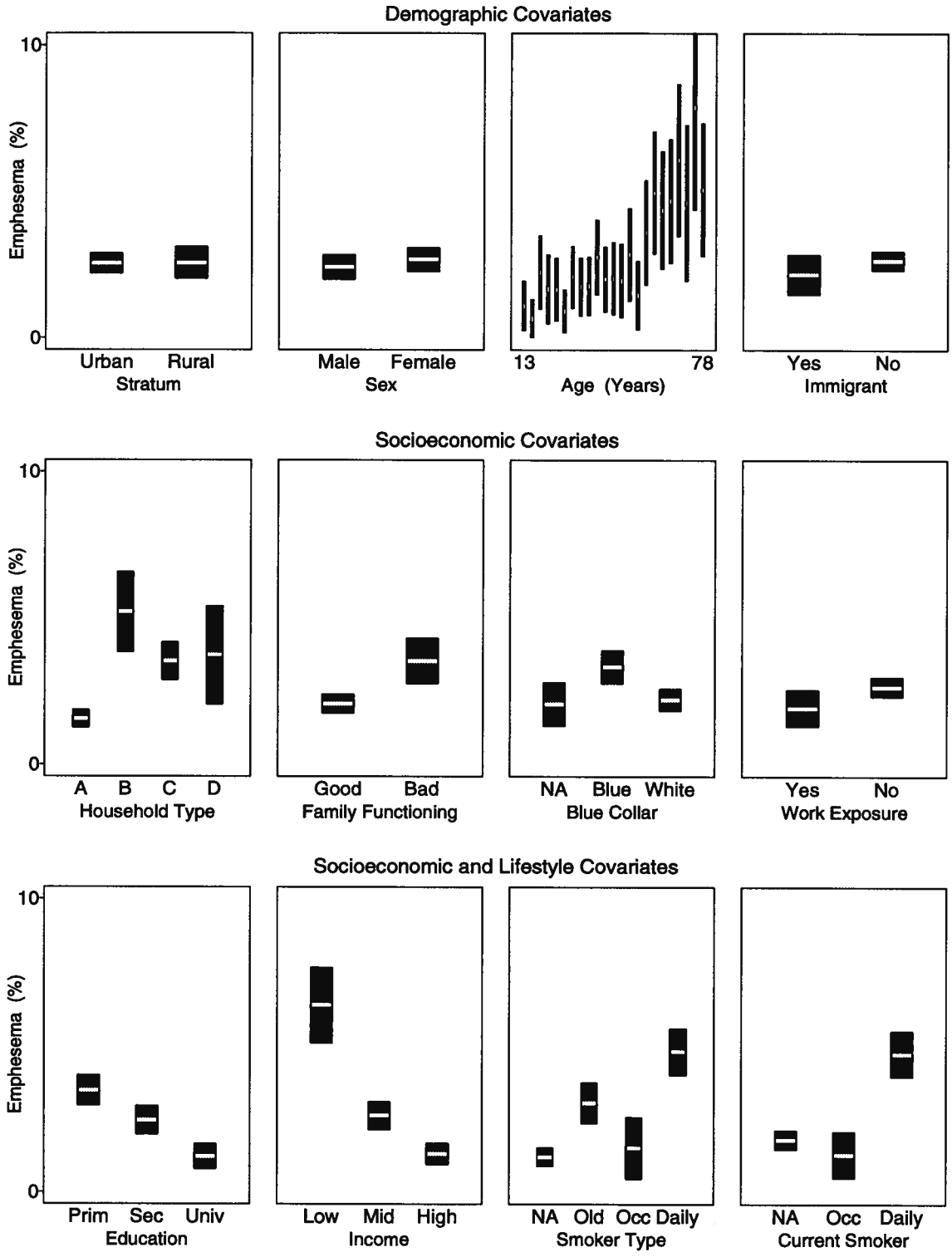

Figure 16: Emphysema prevalence by covariates (I).

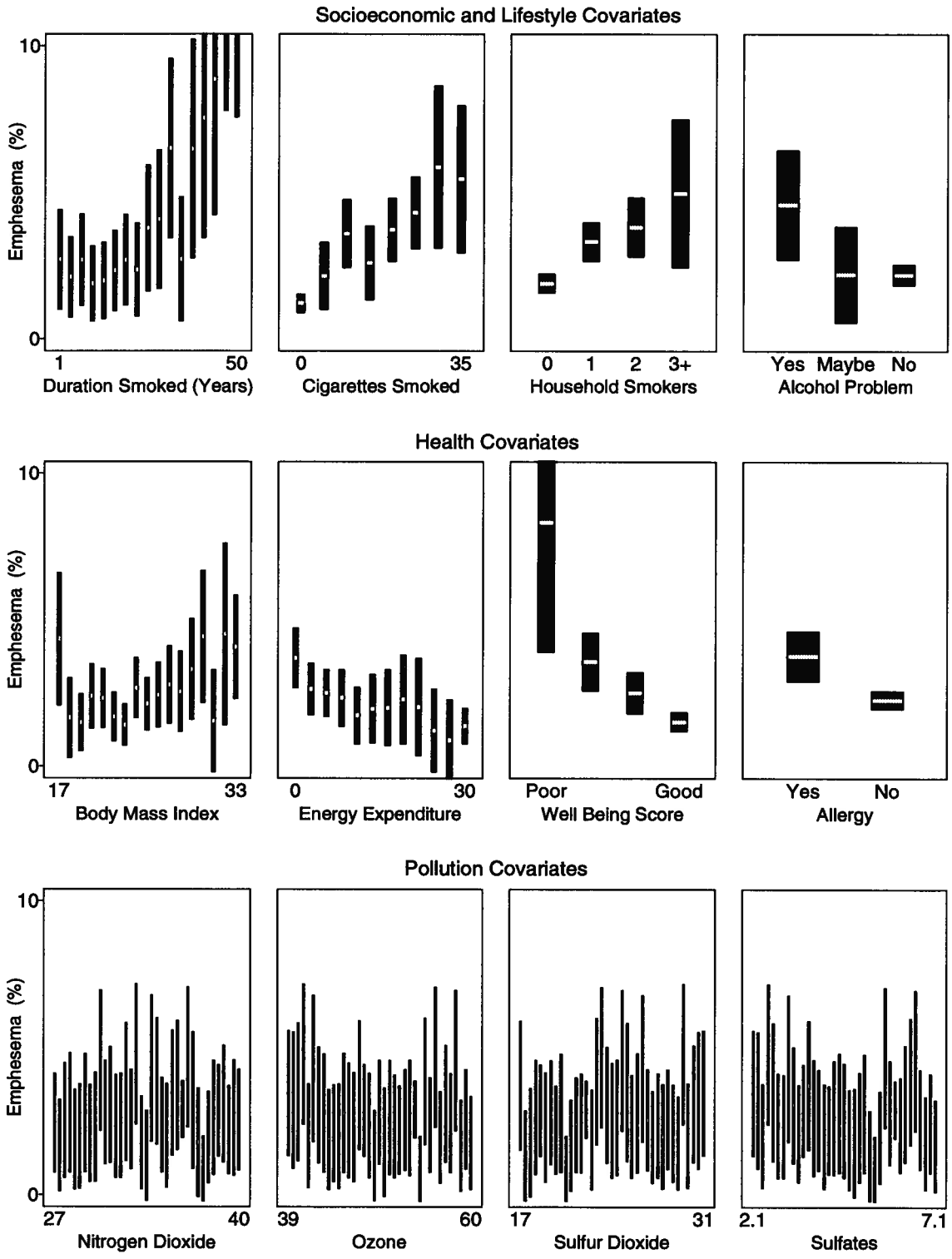

Figure 17: Emphysema prevalence by covariates (II).

6 Methodology

The broad objective of epidemiology is to get at the etiology of disease. Cross sectional data will at best allow for a determination of association between the disease and related factors, casual or otherwise. The specific goal of this study is to identify and assess a number of possibly important explanatory factors contributing to chronic respiratory illness.

Unfortunately the relationship between chronic disease and etiologic agents is muddied by unsure diagnoses, changing environments, multiple causality, changing behavioral habits, the existence of other maladies, measurement error and lack of quality data, to give but a partial list. Each component adds haze to the picture; in statistical jargon, the random component may overshadow the systematic component of the model; in other words, uncertainty threatens to obviate the underlying relationship and hence our understanding of that relationship. Statistical methods address issues of uncertainty and are therefore appropriate in this context.

Besides incorporation of error, the method used ought to be able to examine a number of factors at once. In the epidemiological literature, synergy refers to the phenomenon of two factors which produce a much greater effect in conjunction than if considered separately. Regression methods take care of these concerns. As a bonus, regression provides the possibility of incorporating spatial dependence and other such subtleties.

The Ontario Health Survey sampling design induces a probability structure that is important to address. Should I incorporate inclusion probabilities into the regression model? How can that be done? I begin the chapter by going over the salient differences between finite and infinite population inference.

Classic regression methods assume a continuous response vector. I use responses to the Ontario Health Survey that indicate the presence or absence of chronic respiratory illness. Therefore, before analysing the data, I ought to describe generalized linear models, the extension of classical methods that accommodates binary and binomial response vectors. In this chapter I also describe the model selection strategy and model checking procedures.

6.1 Finite vs. Infinite Population Inference

The Ontario Health Survey data arise from a sample survey of a human population. The sample survey has a complex design which induces inclusion probabilities associated with each element in the sample. Should I incorporate the inclusion probabilities in the analysis? Before the question can be answered I must go over the basics of survey sampling theory.

Finite survey sampling theory gives a way to make inferences from a finite sample, representable as s = $\{1, \ldots, n_s\}$, to a finite universe U = $\{1, \ldots, N\}$. Generally inferential statements require knowledge of the probability structure resulting from the survey design. The setup differs from classical statistical theory in that classical theory assumes an infinite population.

With a finite population the set of all possible samples associated with the design, S = $\{s_1, s_2, \ldots, s_M\}$, is also finite. The probability that a particular sample is chosen, p(s), defines the sample design. Two important probabilities at the element level are

$$\pi_k = P(k^{th} \text{ element sampled}) \qquad \text{and}$$

$$\pi_{kl} = P(k^{th} \text{ and} j^{th} \text{ elements sampled}).$$

If both π_k and π_{kl} can be calculated then the design is said to be measurable. A measurable design allows for estimation of finite population totals and the variance of those estimates. Some papers in the literature deal with estimating finite population regression parameters and their variances. Binder (1983) describes finite parameter estimation in the generalized linear models context. What happens, however, when one wants to consider parameters, such as regression parameters, where the finite population parameters are of little interest?

Classic regression theory posits a linear model ξ relating the continuous response variable to explanatory covariates:

$$y_k = \boldsymbol{x}_k'\boldsymbol{\beta} + \epsilon_k \qquad k = 1, 2, \ldots, n.$$

Further, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independently and identically distributed as N(0, σ^2). To mesh

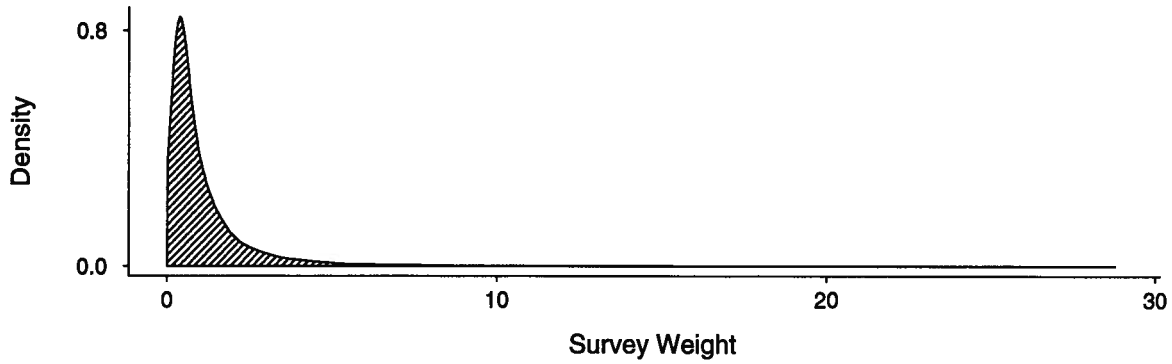

Figure 18: Empirical density of survey weights.

classic regression theory with survey sampling theory, finite survey sampling theorists make use of a construct called the 'superpopulation.'

The finite population is essentially a partial realization of the superpopulation. I can estimate superpopulation parameters in one of two ways: with or without taking the survey design p(s) into account. Here, the parameter $\hat{\boldsymbol{\beta}}_{\mathrm{p(s)}}$ will specify an estimator of $\boldsymbol{\beta}$ which incorporates the inclusion probabilities and $\hat{\boldsymbol{\beta}}$ will refer to an estimator which does not.

If survey weights are more or less the same, a regression analysis incorporating survey weights may differ from an analysis which assume a self weighted design. The survey weight attached to every survey respondent reflects selection probabilities adjusted for nonresponse and age-sex population totals at the PHU level (Ontario Ministry of Health, 1992a). By examining the OHS data, the PHUs with the most extremely weighted respondents are urban. Figure 18 shows the skewness of the weight distribution where the weights have been modified so their expected value is one. Because of the skewness, the question of what to do about survey weights remains.

An ongoing debate is over which of the two estimators is better to use. Särndal (1992) handles the question in the following way. First, he notes $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator

(BLUE). That is, for any conformable constant vector $\boldsymbol{c}$,

$$\mathrm{E}_\xi[\boldsymbol{c}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2 \mid \mathrm{s}, \boldsymbol{X}] \;\leq\; \mathrm{E}_\xi[\boldsymbol{c}'(\hat{\boldsymbol{\beta}}_{other} - \boldsymbol{\beta})^2 \mid \mathrm{s}, \boldsymbol{X}].$$

Under any sampling design $\mathrm{p}(\cdot)$

$$\mathrm{E}_\xi \mathrm{E}_p[\boldsymbol{c}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2 \mid \mathrm{s}, \boldsymbol{X}] \;\leq\; \mathrm{E}_\xi \mathrm{E}_p[\boldsymbol{c}'(\hat{\boldsymbol{\beta}}_{other} - \boldsymbol{\beta})^2 \mid \mathrm{s}, \boldsymbol{X}].$$

By the BLUE criterion, $\hat{\boldsymbol{\beta}}$ is better than $\hat{\boldsymbol{\beta}}_{\mathrm{p(s)}}$. He argues, however, that $\hat{\boldsymbol{\beta}}_{\mathrm{p(s)}}$ may be preferable on the basis that $\hat{\boldsymbol{\beta}}$ is design consistent whereas $\hat{\boldsymbol{\beta}}_{\mathrm{p(s)}}$ is not. Pfeffermann (1993) adds that weights protect against nonignorable sampling designs and misspecification of the model. Either way, a methodology which allows for possible prior weights is desirable.

The OHS documentation suggests the following:

> *The sample weights placed on the individual microdata tape records must be used when producing estimates from the survey data, including ordinary statistical tables.* Otherwise, the estimates derived cannot be considered to be representative of the survey population, and will not correspond to those produced by the Ministry of Health or other users of the data. Users are particularly cautioned about releasing unweighted tables or performing any analysis on unweighted data (Ontario Ministry of Health, 1992b, p. 3).

One way around the controversy is to try the analysis both ways (Fay, 1984). For the marginal distribution of age shown in Figure 19, for example, there are estimates for which there is little practical difference in including or excluding weights.

A nagging difficulty with the superpopulation approach is the assumption of independence between elements. In a realistic application such as the Ontario Health Survey, the units are spatially correlated. Part of the correlation is induced by the clustering and stratification used in the sampling design. Adjustments of the classical χ^2 and likelihood ratio tests are necessary to protect against invalid test statistics (Kumar and Rao, 1984). The assumption of independence normally results in the underestimation of parameter variance.

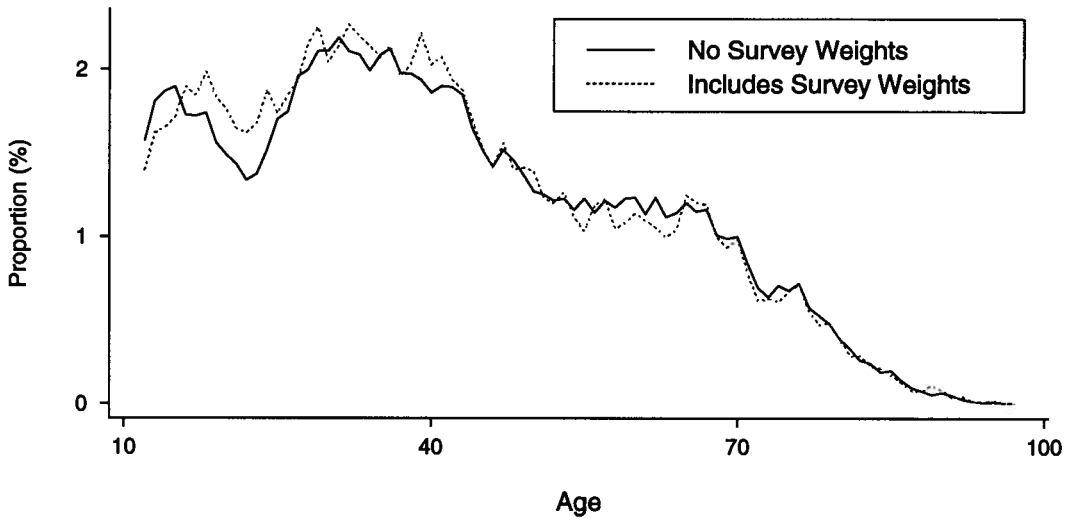

Figure 19: Marginal distribution of age using and ignoring survey weights.

In conclusion, the modelling of survey data introduces a twist into the analysis. A super-population model must be invoked which results in two estimation possibilities: incorporate or ignore inclusion probabilities induced by the survey sampling design.

6.2 Generalized Linear Models Theory

Generalized linear models owe their popularity to the revolution in computer technology. Without a mechanism for solving nonlinear equations using an iterative numerical algorithm, the researcher faces the task of working everything out by hand. Today with the relatively low computational costs, the availability of software like GLIM, SAS and S, the high speed of computers, the use of the generalized model has become quite feasible. Appropriate to such auspicious beginnings, theoreticians have intensely been focusing on generalized linear model problems since the late 1970s; today much of the theory is standardized in "classics" like *Generalized Linear Models* (McCullagh and Nelder, Second Edition, 1989).

Generalized linear models are an extension of classical linear models. The most important change is that the response variable Y no longer needs to be continuous and normally distributed. As long as Y can adequately be described by a member of the exponential family,

generalized linear models theory will apply. The Poisson, binomial, gamma and exponential distributions are examples suggesting the range of possibility for applications.

There are three components to a generalized linear model (GLM):

1. The random component, denoted by the response vector $\boldsymbol{Y}$, is defined by an assumed distributional form. Each element of the response vector belongs to the same exponential family with $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\mu}$.

2. The systematic component is a *linear* function of the explanatory variables $\boldsymbol{x}_j$ where j = 1, 2, $\cdots$, p. The coefficient vector $\boldsymbol{\beta}$ is of great scientific interest since it is used as the criterion for interpreting the importance of covariates as predictors. The systematic component is summarized by $\boldsymbol{\eta} = \sum_1^p \boldsymbol{x}_j\beta_j = \sum_1^n \boldsymbol{x}_i^T\boldsymbol{\beta} = \boldsymbol{X\beta}$.

3. The link function g(·) relates the random and systematic components. Simply put, $\boldsymbol{\eta}$=g($\boldsymbol{\mu}$).

Independence between elements of the response vector is assumed.

The assumed distribution of $\boldsymbol{Y}$ is supposed to adequately model the observed vector $\boldsymbol{y}$. As mentioned, the exponential family includes most well known probability distributions. Properly defined, the exponential family covers all probability density functions and probability mass functions of the form

$$\mathrm{f}_Y(y;\theta,\phi) \;=\; \exp\left\{\frac{y\theta \;-\; b(\theta)}{a(\phi)} \;+\; c(y,\phi)\right\} \tag{2}$$

for given $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. By convention, θ denotes the canonical parameter. When $a(\phi) = \phi/w$, ϕ is called the dispersion parameter where w is a prior known weight associated with the observation y. Until the dispersion parameter comes up explicitly again, I will assume, for the purposes of my argument, that ϕ is known. Questions of estimation then pertain only to the parameter θ.

How can θ be estimated once a distributional assumption about $\boldsymbol{Y}$ is made? If $\mathrm{f}_Y(y;\theta,\phi)$ is regarded as a function of θ alone then one can maximize the likelihood function $L(\theta)$ =

$f_\Theta(\theta; y, \phi)$ to find an estimator by the maximum likelihood approach. As long as the maxima do not occur at the endpoints of the parameter space, the log likelihood

$$l \;=\; \log L(\theta) \;=\; \log f\,(\theta; y, \phi) \;=\; \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \tag{3}$$

will have the same local maxima as $L(\theta)$. The solution for θ is then

$$\left.\frac{\partial l}{\partial \theta}\right|_{\hat{\theta}} \;=\; 0.$$

The log likelihood for $\boldsymbol{Y}$, $l \;=\; \sum \log L(\theta)$, takes the form of a summation over the sample.

There are two important properties of the log likelihood:

$$\mathrm{E}_\theta\left(\frac{\partial \mathrm{l}}{\partial \theta}\right) \;=\; 0 \tag{4}$$

and

$$\mathrm{E}_\theta\left(\frac{\partial^2 \mathrm{l}}{\partial \theta^2}\right) + \mathrm{var}_\theta\left(\frac{\partial \mathrm{l}}{\partial \theta}\right) \;=\; 0. \tag{5}$$

The statistic $\partial l/\partial\theta$ is known as the score. Fishers' information

$$i(\theta) \;=\; \mathrm{var}_\theta\left(\frac{\partial \mathrm{l}}{\partial \theta}\right) \;=\; -\mathrm{E}_\theta\left(\frac{\partial^2 \mathrm{l}}{\partial \theta^2}\right)$$

is derived from (5). Note that the largest value achieved by taking the reciprocal of Fishers' information in the exponential family occurs when the variance is smallest. An analogous inversion takes place when computing standard errors for the estimated coefficient vector $\hat{\boldsymbol{\beta}}$. Thus, the mathematical expression for 'information' meshes nicely with intuition: decreasing uncertainty, i.e. increasing certainty, implies the notion of increasing confidence about the information latent in the data. From (3),

$$\begin{aligned}
\mathrm{E_Y}\left(\frac{\partial \mathrm{l}}{\partial \theta}\right) \;&=\; \mathrm{E_Y}\left(\tfrac{\mathrm{Y}-\mathrm{b}'(\theta)}{\mathrm{a}(\phi)}\right) \;\Longrightarrow\; \mathrm{E(Y)} \;=\; \mathrm{b}'(\theta), \\
\mathrm{E_Y}\left(\frac{\partial \mathrm{l}}{\partial \theta}\right)^2 \;&=\; \mathrm{E_Y}\left(\tfrac{\mathrm{Y}-\mathrm{b}'(\theta)}{\mathrm{a}(\phi)}\right)^2 \;=\; \frac{\mathrm{var(Y)}}{\mathrm{a}^2(\phi)}, \\
\mathrm{E_Y}\left(\frac{\partial^2 \mathrm{l}}{\partial \theta^2}\right) \;&=\; \mathrm{E_Y}\left(\tfrac{-\mathrm{b}''(\theta)}{\mathrm{a}(\theta)}\right) \;=\; \frac{-b''(\theta)}{a(\theta)},
\end{aligned}$$

and

$$\mathrm{E_Y}\left(\frac{\partial^2 \mathrm{l}}{\partial\theta^2}\right) + \mathrm{E_Y}\left(\frac{\mathrm{Y} - \mathrm{b}'(\theta)}{\mathrm{a}(\phi)}\right)^2 = 0 \quad \Longrightarrow \quad \mathrm{var(Y)} = \mathrm{b}''(\theta)\,\mathrm{a}(\phi).$$

A couple of remarks are in order. Assume $a(\phi)$ is of the form $a(\phi) = \phi/w$ where w is a known prior weight. First, E(Y) does not depend on $a(\phi)$. Second, the variance of Y depends both on the dispersion parameter and the prior weight. One possibility for weights are the survey inclusion probabilities associated with each component of $\boldsymbol{y}$.

So far the discussion has shied away from the coefficient vector $\boldsymbol{\beta}$. Observe, however, that

$$\boldsymbol{\mu} = \mathrm{E(Y)} = \mathrm{b}'(\theta) = \mathrm{g}^{-1}(\eta) = \mathrm{g}^{-1}(\boldsymbol{x}^T\boldsymbol{\beta}). \tag{6}$$

If the maximum likelihood estimate (MLE) of θ can be computed then in principle a similar approach should work for finding a $\boldsymbol{\beta}$estimate. Indeed, an application of the principle leads to the Newton Raphson method (Collett, 1992, Appendix B). Let $\boldsymbol{u}(\boldsymbol{\beta})_{q\times 1}$ be the vector of scores where the i^{th} element is $\partial l(\boldsymbol{\beta})/\partial\beta_i$. The objective is to satisfy the maximum likelihood criterion

$$\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \left.\frac{\partial l}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = \boldsymbol{O}. \tag{7}$$

Unfortunately no closed form solution is possible when g($\boldsymbol{\mu}$) is nonlinear since g$'(\boldsymbol{\mu})$ will not reduce to a constant. In other words, the MLE for $\boldsymbol{\mu}$, a function of $\boldsymbol{\beta}$, has no analytic solution. A numerical solution can be obtained, however, by linearization. The first order Taylor linearization of l about the vector $\boldsymbol{\beta}_0$ is

$$\boldsymbol{u}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{u}(\boldsymbol{\beta}_0) + \boldsymbol{H}(\boldsymbol{\beta}_0)\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \tag{8}$$

where $\boldsymbol{H}(\boldsymbol{\beta}_0)_{q\times q}$ is the Hessian matrix with the $(i,j)^{th}$ element given by $\partial^2 l(\boldsymbol{\beta})/\partial\beta_i\partial\beta_j$. From (7) and (8), $\boldsymbol{u}(\boldsymbol{\beta}_0) + \boldsymbol{H}(\boldsymbol{\beta}_0)\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \approx \boldsymbol{O}$ implies

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \boldsymbol{H}^{-1}(\boldsymbol{\beta}_0)\,\boldsymbol{u}(\boldsymbol{\beta}_0).$$

In more standard notation,

$$\hat{\beta}_{r+1} = \hat{\beta}_r + \boldsymbol{H}^{-1}(\hat{\beta}_r)\, \boldsymbol{u}(\hat{\beta}_r) \tag{9}$$

where $\hat{\beta}_r$ is now used to estimate $\hat{\beta}_{r+1}$. The last equation is more obviously in the form of the iterative Newton Raphson procedure for obtaining maximum likelihood estimates of β.

An alternative to Newton Raphson is Fisher's method of scoring. Here the Hessian $\boldsymbol{H}(\beta)$ in (9) is replaced with the information matrix $\boldsymbol{I}(\beta)$. For the $(i, j)^{th}$ element

$$\frac{\partial^2 l(\beta)}{\partial \beta_i \beta_j} \quad \text{is replaced with} \quad -\mathrm{E}\left\{\frac{\partial^2 l(\beta)}{\partial \beta_i \beta_j}\right\}.$$

An advantage of calculating $\boldsymbol{I}(\beta)$ is that the inverse is the asymptotic covariance matrix of the MLE for β. Although both methods converge to the same $\hat{\beta}$, $\boldsymbol{I}^{-1}(\beta)$ will not necessarily be identical to $\boldsymbol{H}^{-1}(\beta)$ for all distributions of the exponential family nor for all link functions.

In conclusion, standardized methodology exists for non-normal data. The likelihood function plays a central role, making coefficient estimation pretty straightforward. The method can handle subtleties like unequal weighting. As important, the methodology has been implemented into major software packages.

6.3 Modelling Binary Data

Now that I have sketched the important components of generalized linear model theory I will focus on the models specifically related to the OHS data. The micro level data gives the presence or absence of chronic respiratory disease for every sampled individual. Two approaches to analysis suggest themselves. One is to go through with a binary variable analysis while the other is to analyse by first aggregating the data and then work with the resulting binomial observations. In this section I will delve into the details of the likelihood equation, link function and parameter estimation for the binary case. The notation I will use is given in Table 7.

Symbol	Description
n	Sample size
π	Probability of disease occurrence
$\boldsymbol{\pi}$	Probability vector
$\hat{\boldsymbol{\pi}}$	Maximum likelihood estimate of $\boldsymbol{\pi}$
Y	Random variable distributed as $\mathrm{B}(1, \pi)$
$\boldsymbol{Y}$	Random vector of binary observations
y	A realization of the random variable Y
w	A prior weight attached to the random variable Y
$\boldsymbol{y}$	A realization of the random vector $\boldsymbol{Y}$
$\boldsymbol{\beta}$	Coefficient vector $(\beta_1, \beta_2, \cdots, \beta_J)^T$
$\hat{\boldsymbol{\beta}}$	Maximum likelihood estimate of $\boldsymbol{\beta}$

Table 7: Definition of symbols used for the modelling of binary data.

A binary variable can take on one of two values:

$$Y = \begin{cases} 1 & \text{when chronic respiratory illness exists,} \\ 0 & \text{otherwise.} \end{cases}$$

The probability $\pi = P(Y = 1)$ is known as the Bernoulli probability of success and results in the probability mass function

$$\mathrm{f}_Y(y; \pi) = \pi^y (1 - \pi)^{1-y} = \left(\frac{\pi}{1 - \pi}\right)^y (1 - \pi).$$

When a weight w_i is attached to each y_i the joint probability mass function appears as

$$\mathrm{f}_Y(\boldsymbol{y}; \boldsymbol{\pi}) = \prod_i \mathrm{f}_Y(w_i, y_i; \pi_i) \tag{10}$$

$$= \exp\left\{\sum_{i=1}^{n} w_i \left[y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + \log(1 - \pi_i)\right]\right\} \tag{11}$$

where $\sum w_i = 1$. The distribution in (11) is a member of the exponential family given in (2) upon setting $\theta = \log[p(1 - p)^{-1}]$. More precisely,

$$\mathrm{f}^*_Y(y_i; \theta_i) = \exp\left\{w_i \left[y_i \theta_i - \log(1 + e^{\theta_i})\right]\right\} \tag{12}$$

because $\log(1 + e^{\theta}) = \log[1 + \pi(1 - \pi)^{-1}] = -\log(1 - \pi)$.

The algorithm for estimating $\boldsymbol{\beta}$ reduces to an iterative weighted least squares algorithm. I am going to give a detailed exposition for this case. Similar arguments can be found in McCullagh and Nelder (1989) and Collett (1991). The outline is simple: only the score function $\boldsymbol{u}(\hat{\boldsymbol{\beta}}_r)$ and Fishers' information matrix $\boldsymbol{I}(\hat{\boldsymbol{\beta}}_r)$ are needed to derive $\hat{\boldsymbol{\beta}}_{r+1}$.

With binary data, the log likelihood is

$$l = \sum w_i \left[y_i \theta_i - \log(1 + e^{\theta_i}) \right].$$

Using the chain rule to derive the score function for β_j,

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial l}{\partial \theta} \frac{\partial \theta}{\partial \pi} \frac{\partial \pi}{\partial \eta} \frac{\partial \eta}{\partial \beta_j} \\ &= \sum w_i \left(y_i - \frac{e^{\theta_i}}{1 + e^{\theta_i}} \right) \left(\frac{1}{\pi_i(1 - \pi_i)} \right) \frac{1}{g'(\eta_i)} x_{ij} \\ &= \sum_i \left(\frac{w_i(y_i - \pi_i)}{\pi_i(1 - \pi_i) g'(\pi_i)} \right) x_{ij}. \end{aligned}$$

For Fishers' information first note that $\mathrm{E}(y_i - \pi_i)(y_{i'} - \pi_{i'}) = 0 \quad \forall i \neq i'$ by the assumed independence between observations. When $i = i'$, $\mathrm{E}(y_i - \pi_i)^2 = \pi_i(1 - \pi_i)$. Thus,

$$\begin{aligned} -\mathrm{E}_Y \left\{ \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} \right\} &= \mathrm{E}_Y \left\{ \frac{\partial l}{\partial \beta_j} \cdot \frac{\partial l}{\partial \beta_k} \right\} \\ &= \mathrm{E}_Y \left\{ \sum_i \frac{w_i(y_i - \pi_i)}{\pi_i(1 - \pi_i) g'(\pi_i)} x_{ij} \sum_{i'} \frac{w_{i'}(y_{i'} - \pi_{i'})}{\pi_{i'}(1 - \pi_{i'}) g'(\pi_{i'})} x_{i'k} \right\} \\ &= \sum_i \frac{w_i^2}{\pi_i(1 - \pi_i)[g'(\pi_i)]^2} x_{ij} x_{ik}. \end{aligned}$$

Grouping terms helps reduce the estimation procedure to iterative weighted least squares. Let the multiplier of $x_{ij} x_{ik}$,

$$W_i(\pi_i) = \frac{w_i^2}{\pi_i(1 - \pi_i)[g'(\pi)]^2},$$

determine the elements of the diagonal $n \times n$ matrix $\boldsymbol{W}$. Clearly $\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{X'WX}$. The score function is simplified by setting $y_i(\pi_i) = [g'(\pi_i)(y_i - \pi_i)] / w_i$. Then $\boldsymbol{u}(\boldsymbol{\beta}) = \boldsymbol{X'W}\boldsymbol{y}(\boldsymbol{\pi})$.

Finally, everything can be combined. Since the MLE $\hat{\boldsymbol{\pi}}$ depends on the estimated value of $\boldsymbol{\beta}$ and vice versa, all terms dependent on $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$ get a subscript r:

$$\hat{\boldsymbol{\beta}}_{r+1} = \hat{\boldsymbol{\beta}}_r + \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}_r)\, \boldsymbol{u}(\hat{\boldsymbol{\beta}}_r)$$

$$
\begin{aligned}
&= \hat{\boldsymbol{\beta}}_r + (\boldsymbol{X}'\boldsymbol{W}_r\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_r\boldsymbol{y}_r(\hat{\boldsymbol{\pi}}_r) \\
&= (\boldsymbol{X}'\boldsymbol{W}_r\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}_r\boldsymbol{X})\hat{\boldsymbol{\beta}}_r + (\boldsymbol{X}'\boldsymbol{W}_r\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_r\boldsymbol{y}_r(\hat{\boldsymbol{\pi}}_r) \\
&= (\boldsymbol{X}'\boldsymbol{W}_r\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}_r[\boldsymbol{X}\hat{\boldsymbol{\beta}}_r + \boldsymbol{y}_r(\hat{\boldsymbol{\pi}}_r)].
\end{aligned}
$$

This is exactly the form of iterative least squares regression on $[\boldsymbol{X}\hat{\boldsymbol{\beta}}_r+\boldsymbol{y}_r(\hat{\boldsymbol{\pi}}_r)]$.

To this point I have made no mention of the specific form of the link function. The logit, probit and complementary log link functions are the three most commonly used for binary and binomial data. Recall that

$$
\pi = \mathrm{E(Y)} = \mathrm{b}'(\theta) = \frac{\mathrm{e}^{\theta}}{1+\mathrm{e}^{\theta}} = \mathrm{g}(\eta).
$$

The advantage of employing one of the three links lies in their range, $(0,1) \in \mathcal{R}$. An unrestricted range could mean, for a given η_0, that $g(\eta_0) < 0$ or $g(\eta_0) > 1$ even though $0 \leq \pi \leq 1$. Since $\eta = g(\pi) = logit(\pi) = \theta$, the logit link is also known as the canonical link. The probit link is defined as $g(\pi) = \Phi^{-1}(\eta)$ where $\Phi^{-1}(\cdot)$ is the inverse cumulative density function for the standard normal distribution and the complementary log function is given as $g(\pi) = \log\{-\log(1-\pi)\}$. I will generally use the logit link for analysis due to its interpretability as the log odds ratio of success.

7 Asthma Analysis

In this chapter we approach model building by dividing the covariates into groups and perform separate arcsine analyses. Besides allowing for standard diagnostic checking, each analysis provides us ultimately with a subset of covariates useful in constructing a 'final' model. This is an alternative to the more commonly used stepwise regression technique.

The arcsine analysis does not allow for the incorporation of survey weights. We therefore opt to use generalized linear models when we add the pollution terms.

7.1 Arcsine Analysis

From Figures 14 and 15 we note the lack of differences for all of the covariates except allergy. Allergy, however, is a questionable covariate since we would hesitate to say allergy causes asthma. Without at least one significant term in a model, asthma will remain unexplained by all covariates including the pollution terms.

With the study variables split into five coherent groupings, an arcsine analysis is possible. This approach transforms the response variable so that it is approximately normally distributed. Then a "classical" regression analysis is possible. The advantage of this approach comes from being able to apply standard diagnostic procedures for model checking. Survey weights will be ignored in this part of the analysis.

If $Y \sim \mathcal{B}(n, \pi)$ is a binomial random variable then the normal distribution $\mathcal{N}(n\pi,\, n\pi[1-\pi])$ approximates Y well if the sample size is large enough. Normal regression theory requires, however, that the response variable has a constant variance, i.e. $Y \sim \mathcal{N}(\mathrm{E}(Y),\, \sigma^2)$, rather than one dependent on π. The arcsine transformation is a transformation which obviates the importance of the proportion in the variance term.

The arcsine transformation derives from the so-called "delta- method" of classical statistics. Let $\psi(\cdot)$ be a transformation and $p = Y/n$. The first order Taylor series expansion about π is

$$\psi(p) \;\doteq\; \psi(\pi) \;+\; \psi'(\pi)\,(p-\pi).$$

Group	Model	df	r^2
Demographic	sex + age + age^2	116	0.15
Socioeconomic	dust exposure + income	46	0.18
Lifestyle	smoker type + drinker	142	0.18
Health	well being + family functioning	508	0.04

Table 8: The best arcsine models fitted for asthma.

Then approximately,

$$\begin{aligned} \mathtt{E}[\psi(p)] &= \psi(\pi) \qquad \text{and} \\ \mathtt{Var}[\psi(p)] &= \mathtt{E}[\psi'(\pi)\,(p-\pi)]^2 \\ &= [\psi'(\pi)]^2\,\mathtt{Var}(\pi). \end{aligned}$$

Let $\psi(\pi) = \sin^{-1}\sqrt{\pi}$. Since $\psi'(\pi) = (2\sqrt{\pi}\sqrt{1-\pi})^{-1}$,

$$\mathtt{Var}[\psi(p)] = \frac{1}{4\,\pi(1-\pi)}\,\frac{\pi(1-\pi)}{n} = \frac{1}{4n}.$$

Thus, $Y^* \sim \mathcal{N}(\sin^{-1}\sqrt{\pi},\, 1/4n)$ where $Y^* = sin^{-1}(p)$. The linear regression analysis becomes a weighted regression analysis.

Our criterion for the arcsine analysis is the standard t-test for each of the covariates. An unusually low p-value suggests the covariate could be important in the final model. The best models for each of the groupings is given in Table 8. Note that although many of the covariates passed this test for admission to the final model, none of the models does a very good job at fitting the data. At best, the models explain about 20% of the observed variation. This leaves another 80% unexplained!

The tables do not tell the whole story, however. For each of the models, we produced standard diagnostic plots (see Figures 34 to 36 in the Appendix). The first noticeable feature is how the sample spreads unevenly in the n-way table. The leverage for most models looks reasonable when compared to the '2p/n' line. In general, if a leverage point lies greatly above the line the model becomes suspect.

Term	X*	df	P (X* > χ^2_{df})
Age + Age^2	52.4	2	0.000
Income	42.7	2	0.000
Smoker Type	21.2	4	0.000
Sex	12.9	1	0.000
NO_2	7.3	1	0.007
O_3	1.9	1	0.169
⋮	⋮	⋮	⋮

Table 9: Most significant terms in the unweighted logit asthma model.

The real test of a model, however, is provided by the residuals. The quantile-quantile plot of normality shows how well the assumption of normality is met. All of the models fare somewhat poorly in this regard. Caution must be exercised in assessing the model.

The socioeconomic covariates of well being and family functioning score do not, upon second reflection, seem to offer much in terms of interpretability of the model. Thus the final model was built from all covariates but those two. The full model consists therefore of 'age', 'sex', 'dust exposure', 'income', 'drinker' and 'smoker type'.

7.2 Logistic Modeling

For fitting the final models, generalized linear model theory is used. The criterion for covariate inclusion in the model is a goodness of fit test. From theory, models can be compared on the basis of differences in model deviance. Mathematically, if $\hat{L}_c$ is the estimated likelihood of the current model and $\hat{L}_f$ the estimated likelihood for the full model, the deviance is conventionally given as

$$D = -2 \log(\hat{L}_c/\hat{L}_f).$$

Nested models can be compared by focusing on the difference in deviances. The difference is asymptotically distributed as χ^2 with the degrees of freedom being the difference of degrees of freedom between the nested models.

The overall fit for the full model is X* = 120 on 12 degrees of freedom. Once the full model

Term	X*	df	P (X* > χ^2_{df})
Age + Age^2	71.8	2	0.000
Income	25.1	2	0.000
Smoker Type	21.6	4	0.000
Sex	10.4	1	0.001
Drinker	12.7	2	0.002
Work Exposure	4.6	1	0.032
NO_2	1.7	1	0.196
⋮	⋮	⋮	⋮

Table 10: Most significant terms in the weighted logit asthma model.

is fitted, forward and backward updating of the model allows us to test for each term. The unweighted analysis produces Table 9.

Age is the most significant term in the model. The fitted values over the domain of ages covered in the sample show a decreasing trend for asthma with a range of about 3%. The next terms are income, smoker type and sex. The model suggests that having lower income, having been a former smoker and being female is associated with higher asthma prevalence.

The model also suggests that NO_2 is associated with with asthma prevalence. The problem is that its estimated coefficient suggests that NO_2 is negatively related to asthma. That is to say, the more NO_2 in the air, the less asthmatics you would expect to observe; the range of the downward trend in fitted values is about 1% over the interval of NO_2 readings observed in Ontario. Two comments are necessary. First, the models have 50,000 observations from which to fit a maximum likelihood estimate and a spurious result is possible. Second, we can check the model estimates be comparing to models which have incorporated the survey weights.

Logistic regression modeling allows for the inclusion of a weighting structure, i.e. survey weights representing inclusion probabilities. The same analysis was run with the weights, producing Table 10. It is comforting to see that the first four significant terms in the unweighted model show up in the weighted model in the same order. In the weighted case, however, drinking and work exposure eclipse NO_2 as important model variables. In fact NO_2 no longer

appears to be significant. The changes reiterate our suspicions of NO_2 as a truly significant covariate.

In summary, none of the pollution covariates appear to be significantly related to asthma prevalence. NO_2 gave the strongest signal of the pollution measures but had a questionable negative estimated coefficient and was not robust enough to show consistency between weighted and unweighted logistic analyses.

8 Emphysema Analysis

In this chapter we model the cross sectional association of respiratory morbidity and air pollution, using the 1990 Ontario Health Survey data. For convenience, we call the binary response variable "emphysema." However, that term will refer collectively to any of emphysema, chronic cough and/or chronic bronchitis. The etiology of chronic respiratory disease, especially as regards the effect of ambient air pollutants, is currently unknown and hence hotly debated. We hope our analysis will contribute usefully to that debate.

In the first step of our analysis, we select appropriate covariates, to increase the model's sensitivity to the effect of pollution while avoiding collinearity and extreme data spread. We use monthly pollutant data derived from the multivariate spatial interpolation methodology described in the last section. In our first approach, we calculate the average summer and winter six year pollution levels for all four pollutants by Public Health Unit (PHU). The resulting values, unlike those from our second approach described next, resemble the measurements taken at the air monitoring stations; the analysis will lend itself to easy interpretation. In our second conceptually more complex approach, we use principal components to construct a pollution index which favours summer pollution levels and represents a variety of pollutants. We obtain the weighting scheme implicitly by using averages of the eight month winter and the four month summer. This leads to eight estimates per Public Health Unit (PHU). A principal components analysis creates the index which extracts the maximal amount of information in the eight estimates.

In our last step we evaluate the significance of the pollution variables. We base our evaluation on the stepwise addition of pollution variables to the covariate model built in the first step. Using the same steps, we go through with an analysis ignoring weights and then offer a comparison to the analysis incorporating weights.

8.1 Covariate Selection Excluding Pollution

We first construct the 'best' model excluding the pollution terms. From previous studies we know that several factors relate to chronic bronchitis. The disease is more prevalent among older people than younger people, among males than females and among urbanites more than rural dwellers. Social class seems important. In Britain unskilled labourers are five times as likely to have chronic bronchitis as professionals. Smoking and family history also rate as possible determinants (Fry, 1985, Chapter 6).

Beginning with the seemingly most relevant twenty, we must select covariates from the many offered by the Ontario Health Study (OHS) to represent broad population traits. But which covariates are best and how many should be included?

Achieving an aesthetic result and avoiding the problems associated with sparse data and collinearity demand parsimony. In particular, a model with too many terms will spread the data over a table of unduly high dimension, since a relatively small proportion of the population is afflicted with emphysema.

Collinearity arises from the association of prospective covariates. Perfect association between them means the second adds no information not in the first. In ordinary least squares regression, collinearity leaves the coefficient estimators unbiased but of reduced precision. Robinson and Jewell (1991) prove that in logistic regression as well, when two predictive covariates exhibit collinearity, one being correlated with the response, a loss of precision occurs. Unfortunately we cannot avoid the difficulties presented by collinearity since covariates in an observational study cannot be controlled. Instead we have tried to side-step these problems by excluding highly correlated pairs of covariates.

To reduce our computational burden, we began building our model one factor at a time using one fifth of the survey data respondents or about 9,500 records. We incorporated the microdata survey weights into the binary logistic models. This insured unbiasedness for the finite population of the estimating equations used to construct estimates of model coefficients.

Term	X^*	df	$P(X^* > \chi^2_{df})$
Age	77	1	0.000
Household Type	47	3	0.000
Education Level	41	2	0.000
Income	50	2	0.000
Smoker Type	103	4	0.000
Current Smoker Type	74	3	0.000
Duration Smoked	150	2	0.000
Number of Cigarettes Smoked	103	2	0.000
Energy Expenditure	63	3	0.000
Well Being Status	56	4	0.000
Number of Current Household Smokers	23	1	0.000
Allergy	20	1	0.000
Alcohol Screening Test Category	24	4	0.000
Blue Collar Work	12	2	0.002
Family Functioning Status	8	2	0.017
Body Mass Index	8	2	0.020
Sex	3	1	0.100
Immigrant Status	2	1	0.214
Work Exposure	1	1	0.333
Stratum	0	1	0.906

Table 11: One term emphysema models using a fifth of the data.

Table 11 summarizes the results. Except for age, the demographic covariates are inconspicuous in that they fall at the bottom of the list. The last four one term models would be rejected at the $\alpha = 0.05$ level criterion; sixteen of the prospective covariates reduce the standard deviance enough to improve the fit of the model significantly. Clearly a judicious strategy for finding a good model is needed.

The need to distinguish between missing and zero values complicated our stepwise selection procedure. When a variable like 'cigarettes smoked' had missing values we had to include an extra dummy variable in the model. In addition, three covariates, 'well being', 'allergy' and 'family functioning status', were excluded a priori since we judged them to blur the distinction between dependent and independent variables.

Table 12 summarizes the results of applying our stepwise procedure. The table shows that adding the smoking covariate in the second step significantly decreases the importance of the other smoking covariates in the subsequent steps.

We stopped after step six . The covariates selected for our model, before considering

Step	Stepwise added Terms	X*	df	P ($X^* > \chi^2_{df}$)
1	Age	77	1	0.000
	Smoker Type	105	4	0.000
	Current Smoker Type	83	3	0.000
2	Cigarettes Smoked	56	1	0.000
	Household Smokers	44	1	0.000
	Education	34	2	0.000
	Education	26	2	0.000
	Income	23	2	0.000
3	Energy Expenditure	25	3	0.000
	Cigarettes Smoked	21	2	0.000
	Immigrant	5	1	0.026
	Cigarettes Smoked	20	2	0.000
	Energy Expenditure	21	3	0.000
4	Family Type	20	3	0.000
	Income	15	2	0.001
	Drinker Category	14	4	0.007
	Sex	6	1	0.018
	Energy Expenditure	21	3	0.000
	Family Type	21	3	0.000
5	Income	17	2	0.000
	Sex	10	1	0.001
	Drinker Category	14	4	0.009
	Family Type	20	3	0.000
	Income	14	2	0.001
6	Sex	8	1	0.004
	Drinker Category	12	4	0.021
	Duration Smoked	7	2	0.025
	Duration Smoked	8	2	0.017
	Sex	5	1	0.027
7	Income	7	2	0.031
	Drinker Category	11	4	0.032
	Immigrant	3	1	0.104

Table 12: Terms in a stepwise fitting strategy for emphysema using a fifth of the data.

Term	X*	df	P ($X^* > \chi^2_{df}$)
Smoker Type	159.2	4	0.000
Age	97.0	1	0.000
Education	85.2	2	0.000
Family Type	98.0	4	0.000
Cigarettes Smoked	24.0	2	0.000
Energy Expenditure	10.1	3	0.018

Table 13: Goodness of fit for each of the terms in the full emphysema model.

pollution, are: 'age'; 'smoker type'; 'education'; 'cigarettes smoked'; 'energy expenditure'; and 'family type'.

8.2 Evaluating the Effect of Pollution

In this subsection we add the pollution variables to the six term model. To check the significance of each of the 'factors' using *all* of the data, we drop each term and note the increase in the deviance. Table 13 shows the improved sensitivity of the model with four times as much data. Except for 'energy expenditure', all factors are more significant than they were with the reduced dataset. Note that, the categorical covariates show a greater improvement than their continuous counterparts.

We considered the pollution covariates in two ways. For the first we computed summer and winter averages; these estimates were easily added to the model. In the second approach we used principal components to reduce the number of pollution covariates. The first three principal components based on the eight averages explained 80% of the variation.

The rotations for the three principal components are given in Table 14 and should be considered indices of long term pollution exposure. A close look at the loadings allows for a certain degree of interpretation. The first contrasts O_3 and SO_4 against SO_2 and emphasizes summer averages; the second is a recasting of NO_2 with most of the weight given to the summer; and the third is a weighted combination of SO_2 and summer O_3.

We evaluated the pollution covariates by comparing nested models containing the pollution

	Component		
Pollutant	#1	#2	#3
Summer NO_2	-0.03	-0.90	0.01
Winter NO_2	-0.09	-0.43	0.09
Summer O_3	0.70	-0.06	0.60
Winter O_3	0.13	0.05	0.03
Summer SO_2	-0.51	0.07	0.67
Winter SO_2	-0.35	0.05	0.38
Summer SO_4	0.26	0.06	0.18
Winter SO_4	0.17	0.00	0.04

Table 14: Loadings for the first three principal components.

Term	X^*	df	$P(X^* > \chi^2_{df})$
Summer NO_2	4.3	1	0.039
Pollution Component #2	3.2	1	0.073
Winter O_3	2.8	1	0.093
Summer O_3	1.1	1	0.299
Summer SO_4	0.8	1	0.380
Pollution Component #1	0.8	1	0.386
Pollution Component #3	0.5	1	0.493
Winter SO_4	0.4	1	0.554
Summer SO_2	0.0	1	0.920
Winter NO_2	0.0	1	1.000
Winter SO_2	0.0	1	1.000

Table 15: Goodness of fit test for the pollution terms.

term to the full model. The resulting decrease in scaled deviance is shown in Table 15. Only 'summer NO_2' significantly improves model fit at the nominal $\alpha = 0.05$ level. The coefficient estimate is 0.013 with a standard error of 0.006. Considering the number of tests we performed and the strong assumptions used in the modeling (such as independence between respondents), the result is more suggestive than irrefutable fact.

8.3 Comparing Weighted and Unweighted Analyses

Since we used weights with the logistic model, we could naturally ask what the results of an analysis would look like if we ignored weights. If results are similar then we can feel confident that model misspecification is minimal.

The same stepwise procedure used previously but without survey weights produces Table 16. As before, 'age' and 'smoker type' give the strongest signals to the model. 'Cigarettes smoked' is again the third term to enter the model. 'Income', however, rather than 'energy expenditure' is the fourth term and, using a similar criterion to the weighted model, the fifth term doesn't even make it into the model. Table 17 shows the goodness of fit test for the terms in the model before adding pollution covariates. Each of the four terms is highly significant.

Table 18 illustrates the addition of pollution covariates. Upon comparing Table 18 with Table 15, we notice that summer NO_2 appears in both as the first pollutant. Moreover the estimated coefficient values are similar with an estimate of 0.012 and a standard error of 0.006.

In both weighted and unweighted models 'age' and 'smoking type' are strongly associated with emphysema. Although summer NO_2 shows up as borderline significant in both cases, significantly stronger covariates such as 'education' and 'income' appear in one model but not the other. We again conclude with a weak belief in the association between summer NO_2 and emphysema.

Step	Stepwise added Terms	X*	df	P ($X^* > \chi^2_{df}$)
	Age	91.3	1	0.000
	Family Type	46.1	3	0.000
1	Income	40.7	2	0.000
	Smoker Type	76.1	4	0.000
	Current Smoker Type	50.5	3	0.000
	Duration Smoked	145	2	0.000
	Smoker Type	77.8	4	0.000
	Current Smoker Type	60.8	3	0.000
2	Cigarettes Smoked	73.7	2	0.000
	Household Smokers	52.8	1	0.000
	Duration Smoked	19.8	1	0.000
	Cigarettes Smoked	24.3	2	0.000
	Allergy	20.4	1	0.000
3	Income	15.8	2	0.000
	Duration Smoked	9.4	2	0.009
	Drinker Category	13.2	4	0.010
	Income	15.7	2	0.000
	Family Type	10.9	3	0.012
4	Household Smokers	5.3	1	0.022
	Drinker Category	10.8	4	0.029
	Education	6.8	2	0.034
	Household Smokers	6.3	1	0.012
	Drinker Category	9.1	4	0.058
5	Duration Smoked	5.7	2	0.059
	Drinker Indicator	4.9	2	0.086
	Family Type	5.8	3	0.122

Table 16: Stepwise terms for emphysema using a fifth of the data and ignoring weights.

Term	X*	df	P ($X^* > \chi^2_{df}$)
Age	297	1	0.000
Smoker Type	123	4	0.000
Income	118	2	0.000
Cigarettes Smoked	46.3	2	0.000

Table 17: Terms in the full unweighted emphysema model.

Term	X*	df	P ($X^* > \chi^2_{df}$)
Summer NO_2	3.8	1	0.051
Pollution Component #2	3.7	1	0.053
Winter NO_2	3.4	1	0.067
Summer SO_4	1.9	1	0.173
Pollution Component #1	1.1	1	0.286
Winter O_3	1.1	1	0.296
Winter SO_4	1.1	1	0.303
Summer O_3	1.0	1	0.325
Summer SO_2	0.2	1	0.699
Pollution Component #3	0.2	1	0.699
Winter SO_2	0.1	1	0.806

Table 18: Goodness of fit test for the pollution terms in the unweighted analysis.

9 Discussion

In this study we explored the association between between airborne pollutants and chronic respiratory health. Our findings at most suggest that a weak association exists. In this chapter we critically consider the model assumptions and exposure measurement problems.

9.1 Model Assumptions

If the pollution coefficients showed strong significance we would inevitably get caught up in a debate over the reported standard errors. Measurement error and clustering used in survey sampling are two reasons why the standard errors are, perhaps even grossly, underestimated.

We normally underestimate standard errors because we implicitly assume the absence of measurement error. We have reason to believe, however, that measurement error does exist in our study variables, the extent to which is unknown. Breaking down the layers of error helps for gaining an understanding of the phenomenon.

The first problem arises from self reporting. When asked whether they have a chronic cough, respondents introduce subjective assessments of their own condition. We all know of psychosomatic individuals and their converse, obdurate, self denying stoics. More precisely worded questions concerning the nature of the symptoms may minimize interpretive leeway. In Abbey, Petersen, Mills and Beeson (1993) symptom assessment questions incorporated specific time durations. "Did you have symptoms of cough and/or sputum production on most days, for at least three months per year, for two years or more?" is one example of the wording. The Ontario Health Survey, on the other hand, asked if the respondent had "emphysema or chronic bronchitis or persistent cough?".

Besides interpretive difficulties associated with a questionnaire, disease diagnosis can be difficult. Emphysema provides a sterling example: a definitive answer avails itself only after autopsy. The two main diagnostic instruments, providing physiological and radiological measurements, are not failsafe.

Physiological measurements include carbon monoxide diffusion capacity, slope of the volume pressure curve and lung recall at specific lung volumes. At best, they have been shown to be good at detecting severe cases; mild disease is poorly detected. Overall, no physiological measure or combination of measurements can reliably serve for disease determination.

Radiography provides an opportunity to visually look inside the body without using the scalpel. Standard films do not do away with the subjective judgement of the medical practitioner, however. As with physiological measurements, chest radiographs are good for diagnosing severe cases but can only diagnose about have the moderately severe cases. Computed tomography offers an improved image over standard radiography and, as a result, better diagnosis potential. High resolution computed tomography, in particular, appears to correctly resolve 90% of all cases (Snider, 1992, p. 1342).

A question arises, of course: if a trained expert is unable to diagnose emphysema reliably then how much caution should we take in considering a self reported condition? We can enumerate some of the germane factors: the patient's willingness to go to the doctor, the date of the last checkup, the severity of the condition, the equipment available to the physician, the physician's willingness to use the equipment and professional judgement determine to some degree the binary response recorded by the 1990 Ontario Health Survey.

Besides measurement error, clustering usually inflates the real error. In the models we considered we assume independence between observations. The Ontario Health Survey, however, uses cluster sampling in which blocks of households, ostensibly more similar to one another the closer they are, are surveyed at the same time. We ignored this phenomenon because the micro-data we had did not allow us to identify the clusters.

In short we used liberal assumptions and underestimated standard errors associated with coefficient estimates. Converting to p-values, the results which we have shown seem more significant than they are. The reader is therefore encouraged to be cautious about accepting statements of significance at face value.

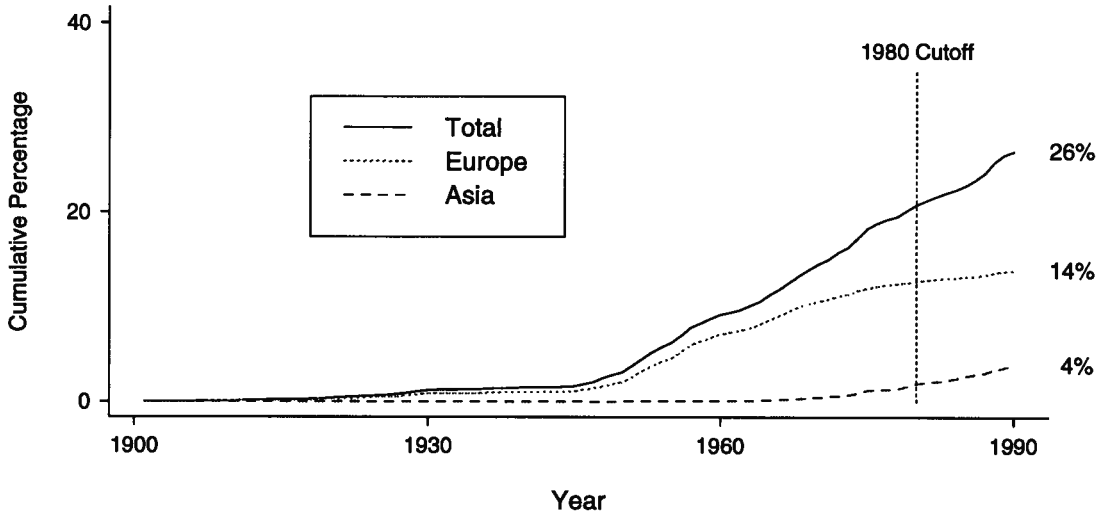

Figure 20: Immigration background of the 1990 residents of Ontario.

9.2 Exposure Measurement Problems

The most difficult task for an epidemiological study of this type is to derive good estimates of long term exposure. Our pollution data may not have been adequate to detect the hypothesized relationship between ambient air pollution and pulmonary disease. We know, for instance, that the study population is anything but fixed. In our study, however, we assumed that people were more or less situated within a Public Health Unit for a time span long enough to make the pollution estimates relevant. Figure 20, estimated from the Ontario Health Survey data itself, shows that at about 25% of the sample immigrated from another country! Other sources report that in the fifteen year span from 1971 to 1986 almost half of all Canadians changed their place of residence every five years (Statistics Canada, 1991, p. 72). Information on the migratory history of the individual's in the sample would have been helpful.

Another difficulty arises from incomplete exposure estimate coverage. In our case we have pollution measurements from 1983 to 1989. We then assume that this time period is similar to earlier intervals. The North American experience, however, leaves that assumption in doubt.

Hoberg (1989) argues that pollution controls have made a difference to observable pollution

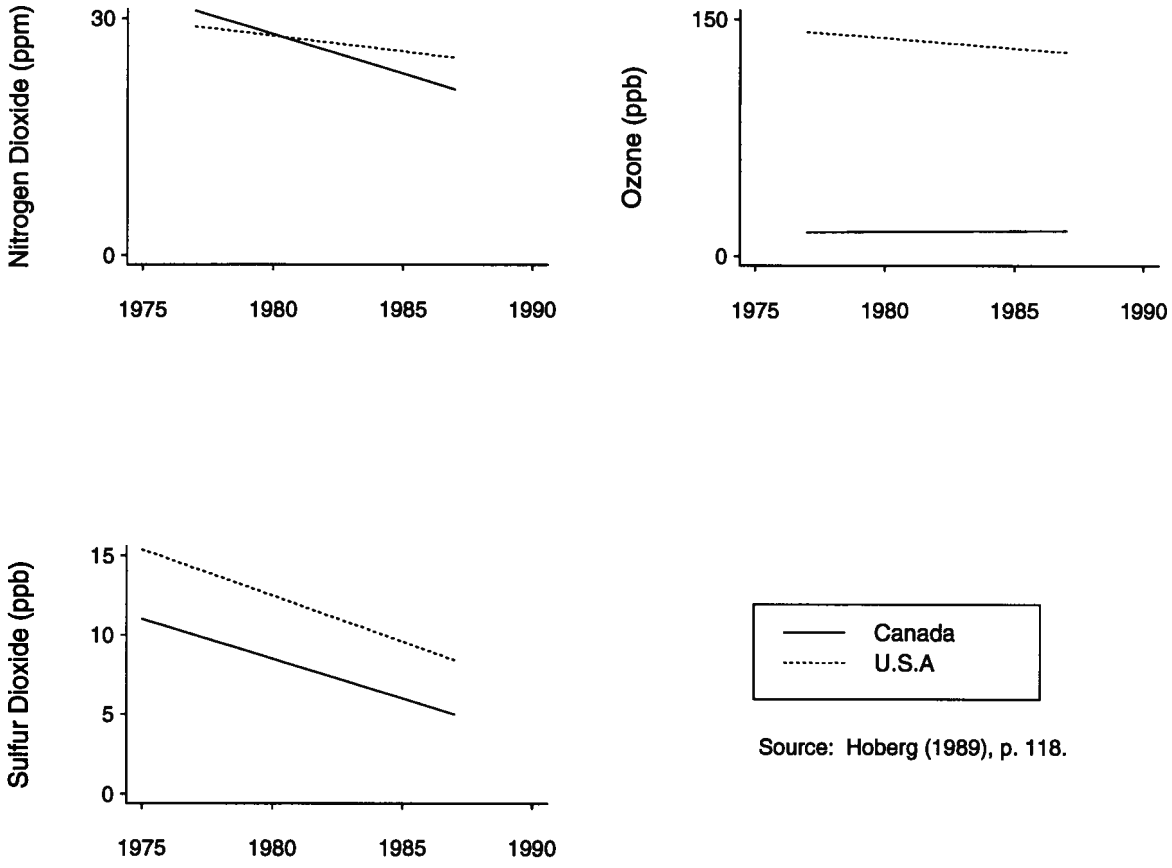

Figure 21: North American air quality trends.

values. Figure 21 shows changes in Canadian and American pollution levels from the mid 1970s to the late 1980s. Though the picture is complex we can generally see a decline in ambient air pollution readings. Changes in the spatial distribution of pollution will also muddy the interpretation of the resulting analysis.

Finally, ambient air pollution could be less important than accumulated levels achieved indoors. Dales et al. (1991a) and Dales, Burnett and Zwanenburg (1991b) identified the importance of indoor molds on the respiratory function of children and adults, respectively. The models I considered for this thesis ignored the potential of household pollutants because of the absence of this kind of data.

9.3 Future Directions

If exposure estimates are often the weak link in epidemiological studies then future efforts must concentrate on improving alternatives for the future. The use of personal monitoring devices offer one such possibility. Silverman et al. (1992) offer a successful application of portable multipollutant samplers. Due to expense, however, only thirty six subjects participated in the study.

If personal monitoring is too costly and fixed station data too crude, a greater emphasis on sophisticated modelling offers an intermediary solution. The Office of Air Quality Planning and Standards, a branch of the U.S. Environmental Protection Agency, simulated people's movements through zones of varying air quality to approximate exposure patterns. The National Ambient Air Quality Standards Exposure Model (NEM) uses ambient air pollution station measurements, activity diaries and population data as its major data sources. Since the early stages of development in 1979, the model has shown promise (Johnson, Capel, Paul and Wijnberg, 1992).

We presume the relationship between airborne pollutants and chronic respiratory disease, if any, is subtle. This study has demonstrated the difficulty of exploring such a relationship through the use of a general purpose survey.

References

[1] ABBEY, David, John MOORE, Floyd PETERSEN and Larry BEESON (1991). "Estimating cumulative ambient concentrations of air pollutants: description and precision of methods used for an epidemiological study." *Archives of Environmental Health.* **46**, 5, 281-287.

[2] ABBEY, David, Floyd PETERSEN, Paul MILLS and Lawrence BEESON (1993). "Long-term ambient concentrations of total suspended particulates, ozone, and sulfur dioxide and respiratory symptoms in a nonsmoking population." *Archives of Environmental Health.* **48**, 1, 33-46.

[3] BATES, David (1992). "Health indices of the adverse effects of air pollution: the question of coherence." *Environmental Research.* **59**, 336-349.

[4] BINDER, David (1983). "On the variances of asymptotically normal estimators from complex surveys." *International Statistical Review.* **51**, 279-292.

[5] BRAIN, Joseph D. (1989). "The susceptible individual: an overview." *Susceptibility to Inhaled Pollutants, ASTM STP 1024.* Mark Utell and Robert Frank, Editors, 3-5.

[6] BRITTON, J. (1992). "Pollution and respiratory morbidity: how much do we accept?" *Thorax.* **47**, 5, 391-392.

[7] BROWN, P.J., LE, Nhu D. and ZIDEK, J.V. (1993). "Multivariate Spatial Interpolation and Exposure to Air Pollutants." *Canadian Journal of Statistics.* To appear.

[8] BURKE, T., H. ANDERSON, N. BEACH, D. COLOME, M. FIRESTONE, F. HUACHMAN, T. MILLER, D. WAGENER, L. ZEISE, and L. TRAN (1992). "Role of exposure databases in risk management." *Archives of Environmental Health.* **47**, 6, 421-429.

[9] CLEMMESEN, Johannes (1993). "Lung cancer from smoking: delays and attitudes, 1912-1965." *American Journal of Industrial Medicine.* **23**, 941-953.

[10] COLLETT, Dave (1991). *Modelling Binary Data.* London: Chapman and Hall.

[11] DALES, Harry ZWANENBURG, Richard BURNETT and Claire FRANKLIN (1991a). "Respiratory health effects of home dampness and molds among Canadian children." *American Journal of Epidemiology.* **134**, 2, 196-203.

[12] DALES, Robert, Richard BURNETT and Harry ZWANENBURG (1991b). "Adverse health effects among adults exposed to home dampness and molds." *American Review of Respiratory Disease.* **143**, 505-509.

[13] DOBSON, Annette J. (1990). *An Introduction to Generalized Linear Models.* London: Chapman and Hall.

[14] DUDDEK, Chris, Nhu LE, Weimin SUN, Richard WHITE, Hubert WONG and James V. ZIDEK (1994). "Assessing the impact of ambient air pollution on hospital admissions using interpolated exposure estimates in both space and time." Final report to Health Canada under DSS contract H4078-3-C059/01-SS.

[15] EVANS, J.S., T. TOSTESON and P.L. KINNEY (1984). "Cross sectional mortality studies and air pollution risk assessment." *Environment International.* **10**, 55-83.

[16] EVANS, Robert (1993). "Less is more: Contrasting styles in health care." In *Canada and the United States: Differences that Count.* Edited by David Thomas. Peterborough, Ont: Broadview Press. 21-41.

[17] FAY, Robert E. (1984). "Application of linear and log-linear models to data from complex surveys." *Suvery Methodology.* **10**, 1, 82-96.

[18] FRY, John (1985). *Common Diseases: Their Nature, Incidence and Care.* Fourth Edition. Lancaster: MTP Press Limited.

[19] GIBSON, Robert (1990). "Out of control and beyond understanding: Acid rain as a political dilemma." In *Managing Leviathan: Environmental Politics and the Administrative*

State. Edited by Robert Paehlke and Douglas Torgerson. Peterborough, Ont: Broadview Press. 243-282.

[20] GOMEZ, Stephen, Robert PARKER, James DOSMAN and Helen McDUFFIE (1992). "Respiratory health effects of alkali dust in residents near desiccated Old Wives Lake." *Archives of Environmental Health.* **47**, 5, 364-369.

[21] HACKNEY, Jack, William LINN, Edward AVOL, Deborah SHAMOO, Karen ANDERSON, Joseph SOLOMON, David LITTLE and Ru-Chuan PENG (1992). "Exposures of older adults with chronic respiratory illness to nitrogen dioxide." *American Review of Respiratory Disease.* **146**, 1480-1486.

[22] HOBERG, George (1993). "Comparing Canadian performance in environmental policy." In *Canada and the United States: Differences that Count.* Edited by David Thomas. Peterborough, Ont: Broadview Press. 101-124.

[23] INSTITUTE FOR HEALTH CARE FACILITIES OF THE FUTURE (1990). *Future Health: A View of the Regional Trends.* Ottawa.

[24] JOHNSON, Ted, Jim CAPEL, Roy PAUL and Luke WIJNBERG (1992). "Estimation of carbon monoxide exposures and associated carboxyhemoglobin levels in Denver residents using a probabilistic version of NEM." Final report to U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, contract number 68-D0-0062.

[25] KUMAR, S. and J.N.K. Rao (1984). "Logistic regression analysis of labour force survey data." *Survey Methodology.* **10**, 1, 62-76.

[26] LEBOWITZ, Michael (1991). "Populations at risk: addressing health effects due to complex mixtures with a focus on respiratory effects." *Environmental Health Perspectives.* **95**, 35-38.

[27] MATANOSKI, G., S. SELEVAN, G. AKLAND, R. BORNSCHEIN, D. DOCKERY, L. EDMONDS, A. GREIFE, M. MEHLMAN, G. SHAW and E. ELLIOTT (1992). "Role of Exposure Databases in Epidemiology." *Archives of Environmental Health.* **47**, 6, 439-446.

[28] McCULLAGH, P. and J.A. NELDER (1989). *Generalized Linear Models.* Cambridge: Chapman and Hall.

[29] McDONNELL, Howard KEHRL, Said ABDUL-SALAAM, Philip IVES, Lawrence FOLINSBEE, Robert DEVLIN, John O'NEIL and Donald HORSTMAN (1991). "Respiratory response of humans exposed to low levels of ozone for 6.6 hours." *Archives of Environmental Health.* **46**, 3, 145-150.

[30] McDONNELL, William, Keith MULLER, Philip BROMBURG and Carl SHY (1993). "Predictors of individual differences in acute response to ozone exposure." *American Review of Respiratory Disease.* **147**, 818-825.

[31] ONTARIO MINISTRY OF HEALTH (1992a). "Ontario Health Survey 1990." User's guide volume 1.

[32] ONTARIO MINISTRY OF HEALTH (1992b). "Ontario Health Survey 1990." User's guide volume 2."

[33] ÖZKAYNAK, Halûk and George THURSTON (1987). "Associations between 1980 U.S. mortality rates and alternative measures of airborne particle concentration." *Risk Analysis.* **7**, 4, 449-461.

[34] PFEFFERMANN, Danny (1993). "The role of sampling weights when modeling survey data." *International Statistical Review.* **61**, 2, 317-338.

[35] PURTILO, D. and R. PURTILO (1989). *Survey of Human Diseases.* Second Edition. Boston: Little Brown and Company.

[36] RAPPAPORT, S.M., (1993). "Threshold limit values, permissible exposure limits and feasibility: the bases for exposure limits in the United States." *American Journal of Industrial Medicine*, **23**, 683-694.

[37] ROBINSON, L. and N. JEWELL (1991). "Some surprising results about covariate adjustment in logistic regression models." *International Statistical Review*, **59**, 227-240.

[38] ROBINSON, J. and D. PAXMAN, (1992). "The role of threshold limit values in U.S. air pollution policy." *American Journal of Industrial Medicine*, **21**, 383-396.

[39] SÄRNDAL, Carl-Erik, Bengt SWENSSON and Jan WRETMAN (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

[40] SELZER, M., L VAN ROOIJEN (1975). "A self administered Short Michigan Alcohol Screening Test (SMAST)." *Journal of Studies on Alcohol* **36**, 1, 117-126.

[41] SEIXAS, Noah, Thomas ROBINS and Mark BECKER (1993). "A novel approach to the characterization of cumulative exposure for the study of chronic occupational disease." *American Journal of Epidemiology.* **137**, 4, 463-471.

[42] SEILER, Tamara (1993). "Melting Pot and Mosaic: Images and Realities." In *Canada and the United States: Differences that Count.* Edited by David Thomas. Peterborough, Ont: Broadview Press. 303-325.

[43] SENTHILSELVAN, A., Yue CHEN and James DOSMAN (1993). "Predictors of asthma and wheezing in adults: grain farming, sex and smoking." *American Review of Respiratory Disease.* **148**, 667-670.

[44] SEXTON, K., D. WAGENER and J. LYBARGER (1992). "Estimating human exposures to environmental pollutants: availability and utility of existing databases." *Archives of Environmental Health*, **47**, 6, 398-407.

[45] SILVERMAN, F., H.R. HOSEIN, P. COREY, S. HOLTON and S.M. TARLO (1992). "Effects of particulate matter exposure and medication use on asthmatics." *Archives of Environmental Health.* **47**, 1, 51-56.

[46] SNIDER, Gordon (1992). "Emphysema: the first two centuries and beyond." *American Review of Respiratory Disease.* **146**, 1334-1344.

[47] STATISTICS CANADA (1983). *Historical Statistics of Canada.* Second Edition. Edited by F. H. Leacy and M. C. Urquhart. Ottawa: Supply and Services Canada.

[48] STATISTICS CANADA (1991). *Canada Year Book.* Ottawa: Supply and Services Canada.

[49] STEINBERG, Stephen (1981). *The Ethnic Myth.* Boston: Beacon Press.

[50] VALANIS, Barbara (1992). *Epidemiology in Nursing and Health Care.* Second Edition. Toronto: Prentice Hall.

[51] XIPING, Xu, Douglas DOCKERY and Lihua WANG (1991). "Effects of air pollution on adult pulmonary function." *Archives of Environmental Health.* **46**, 4, 198-206.

Appendix: Figures

I decided to place some of the figures in this appendix in the hope of avoiding unnecessary distraction in the text. The figures are ordered into the following logical groupings:

OHS Nonresponse

Figure 22: Ordered summary of covariate nonresponse.

Pollution Measurements and Estimates

Figure 23: NO_2 station measurements;
Figure 24: O_3 station measurements;
Figure 25: SO_2 station measurements;
Figure 26: SO_4 station measurements;
Figure 27: Strongest relationships between pollutants;
Figure 28: Estimated six year estimates for all pollutants;
Figure 29: Comparison of pollution measurments to estimates;
Figure 30: NO_2 PHU estimates;
Figure 31: O_3 PHU estimates;
Figure 32: SO_2 PHU estimates; and
Figure 33: SO_4 PHU estimates.

Asthma Arcsine Analysis Diagnositcs

Figure 34: Demographic model;
Figure 35: Soccioeconomic model;
Figure 36: Lifestyle model;

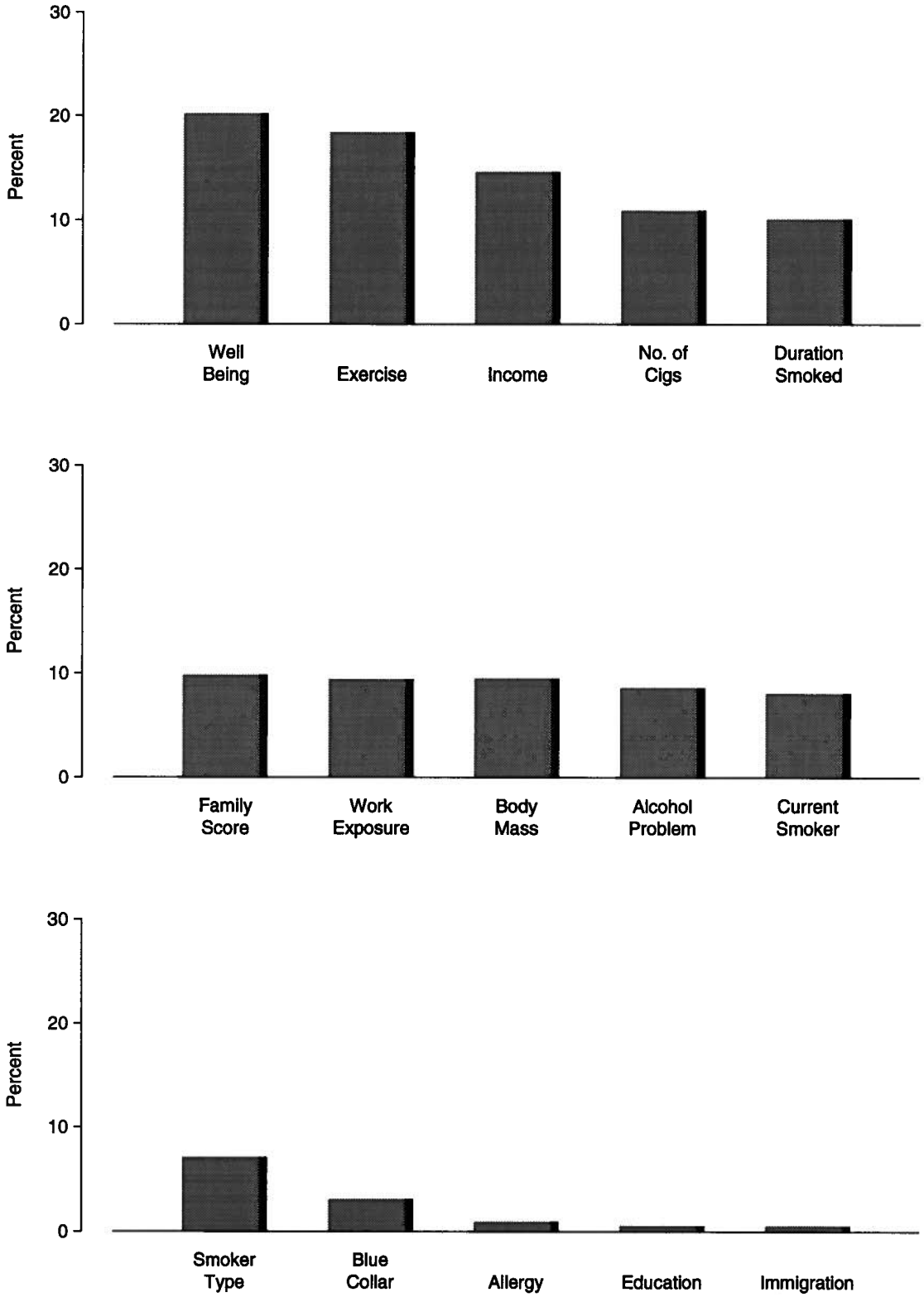

Figure 22: Ordered summary of study covariate nonresponse.

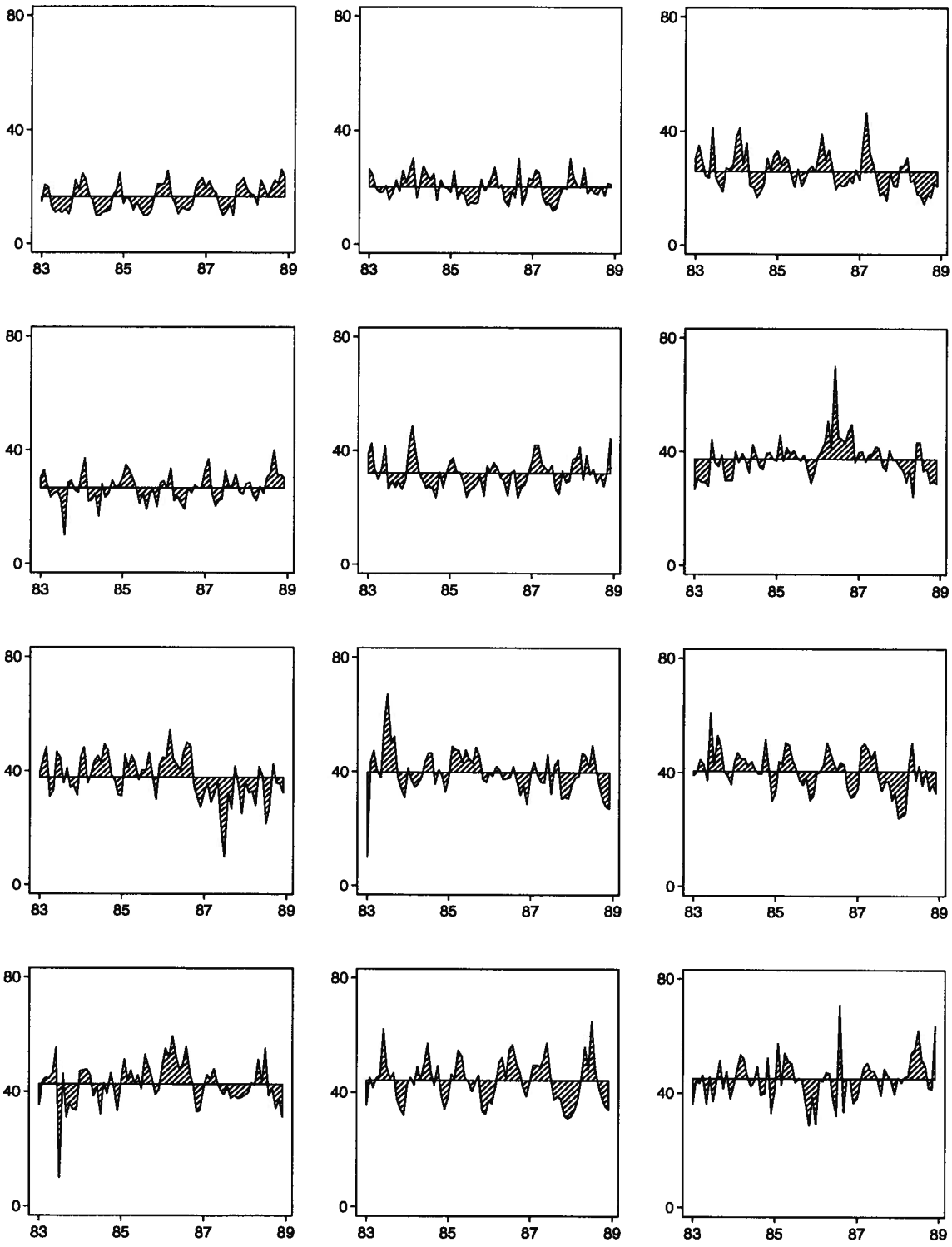

Figure 23: NO_2 readings for twelve stations ordered by increasing station mean ($\mu g/m^3$).

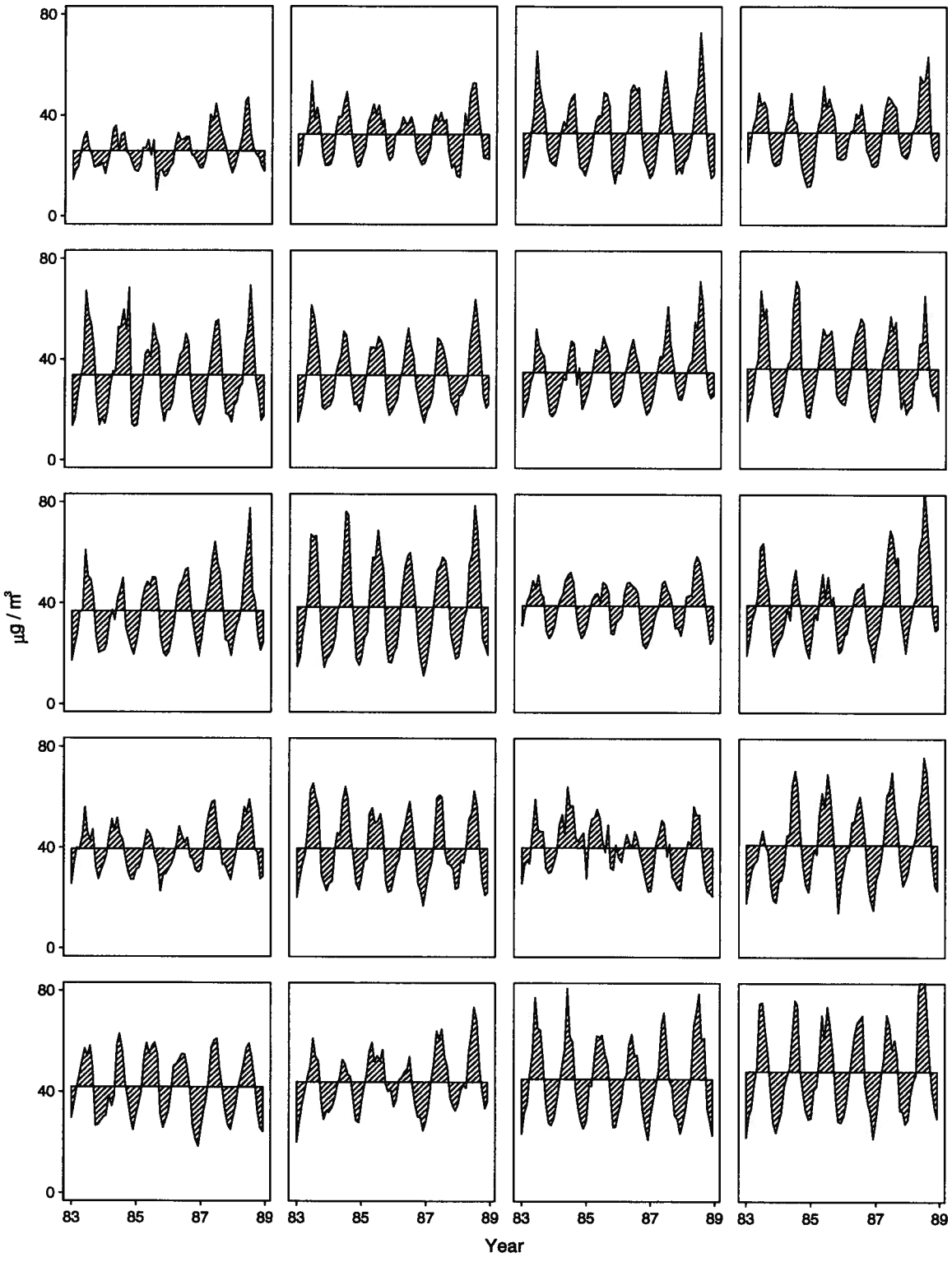

Figure 24: O_3 readings for twenty out of twenty one stations ($\mu g/m^3$).

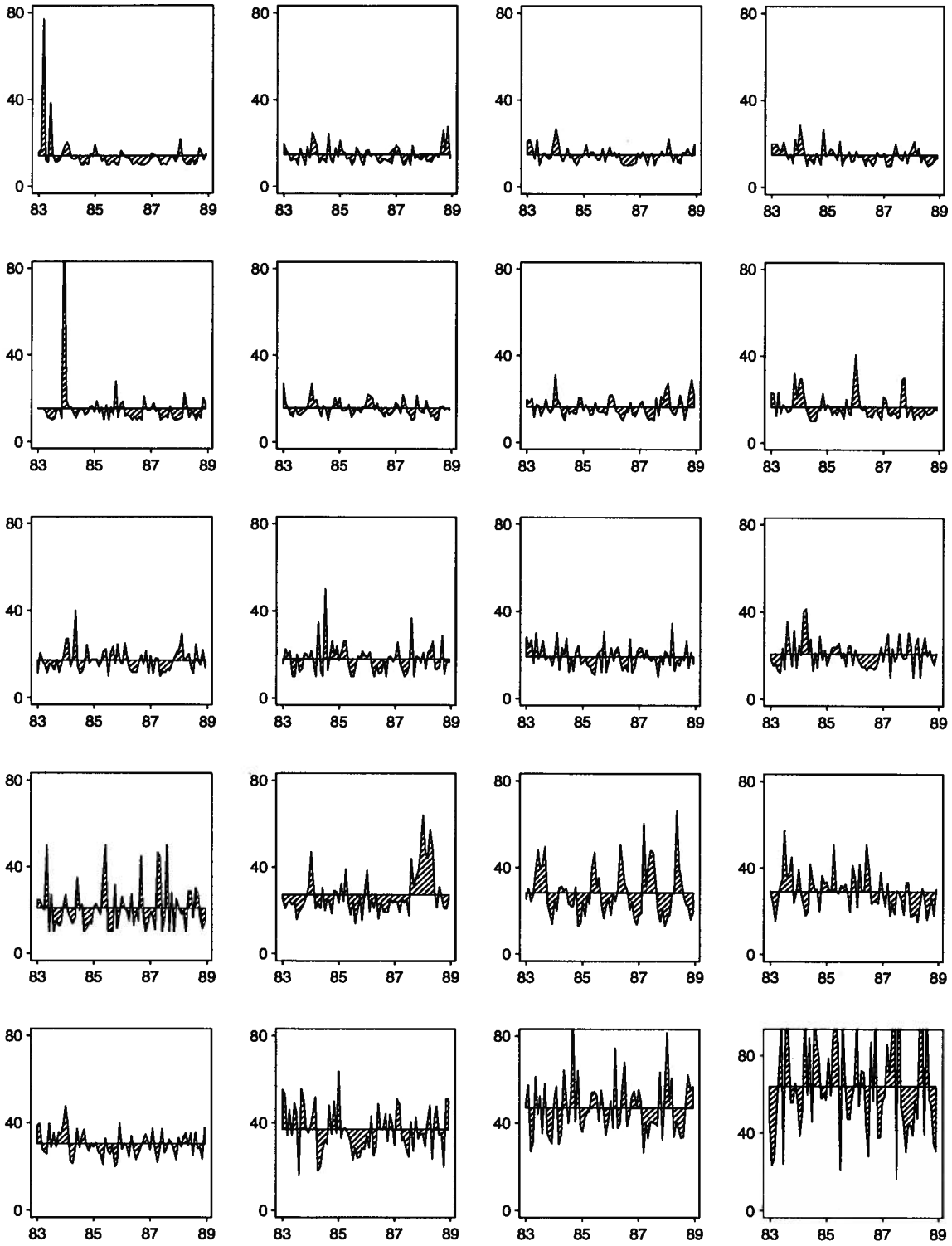

Figure 25: SO_2 readings for all twenty stations ($\mu g/m^3$).

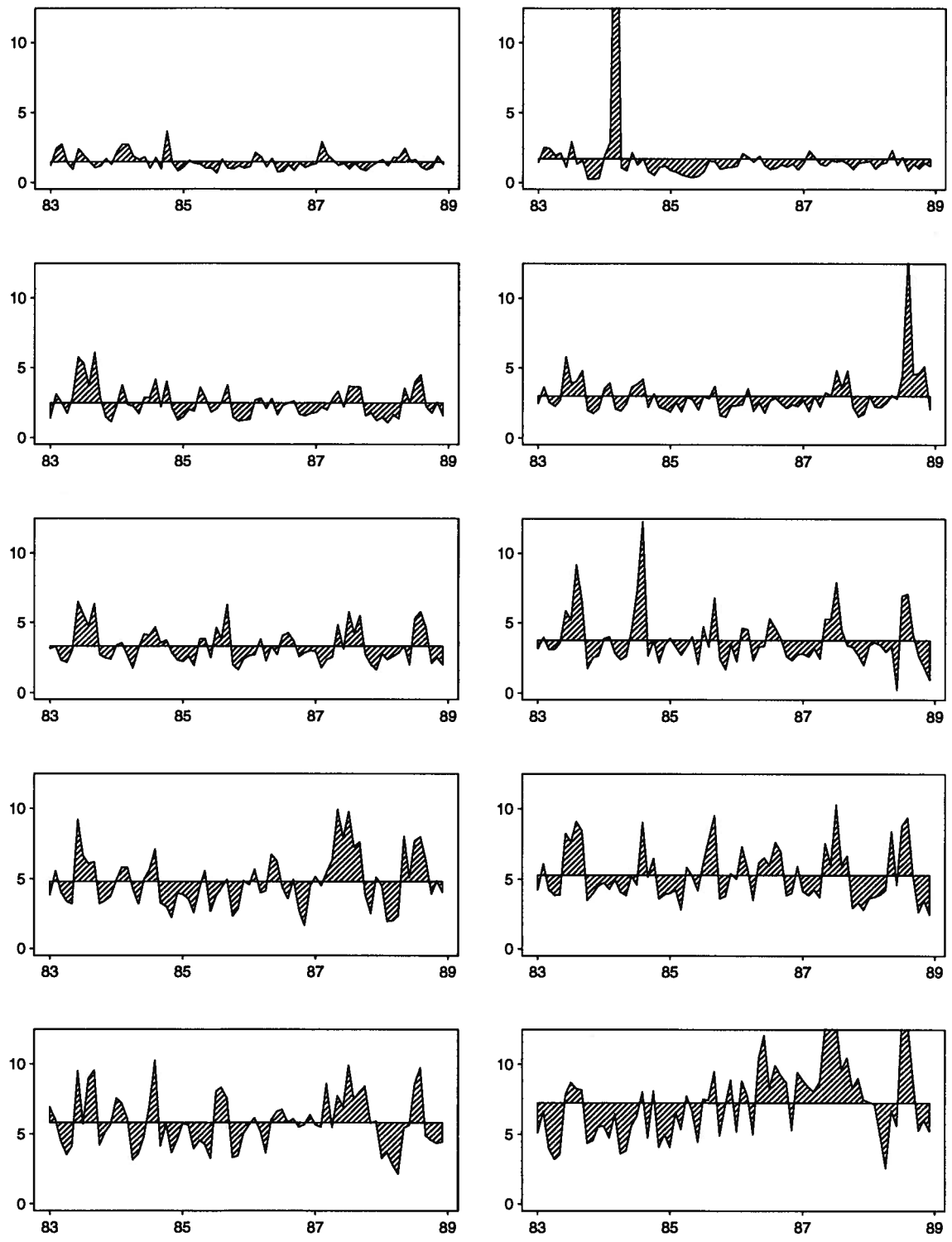

Figure 26: SO_4 readings for all ten stations ($\mu g/m^3$).

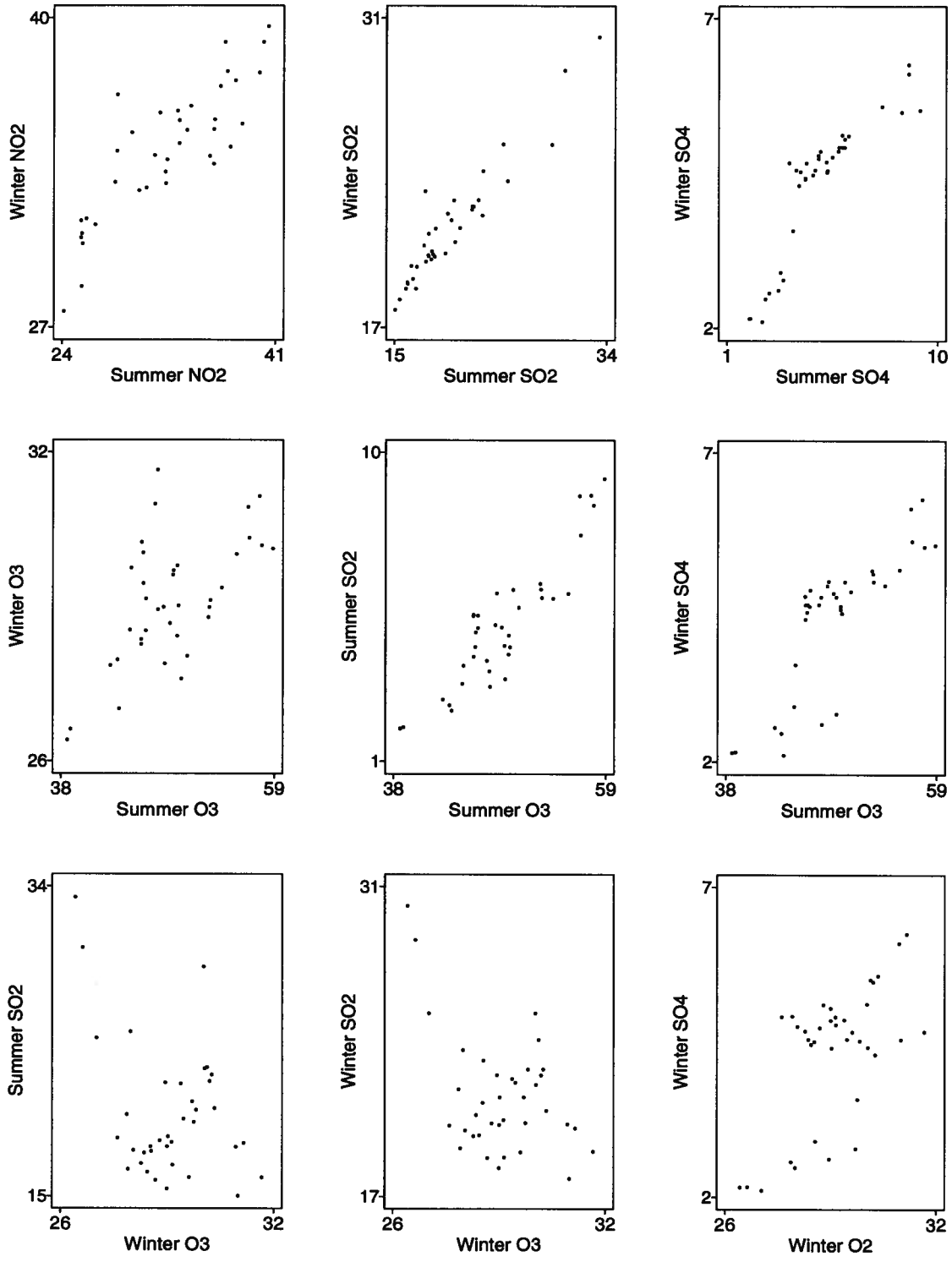

Figure 27: Strongest scatterplot relationships between pollution estimates.

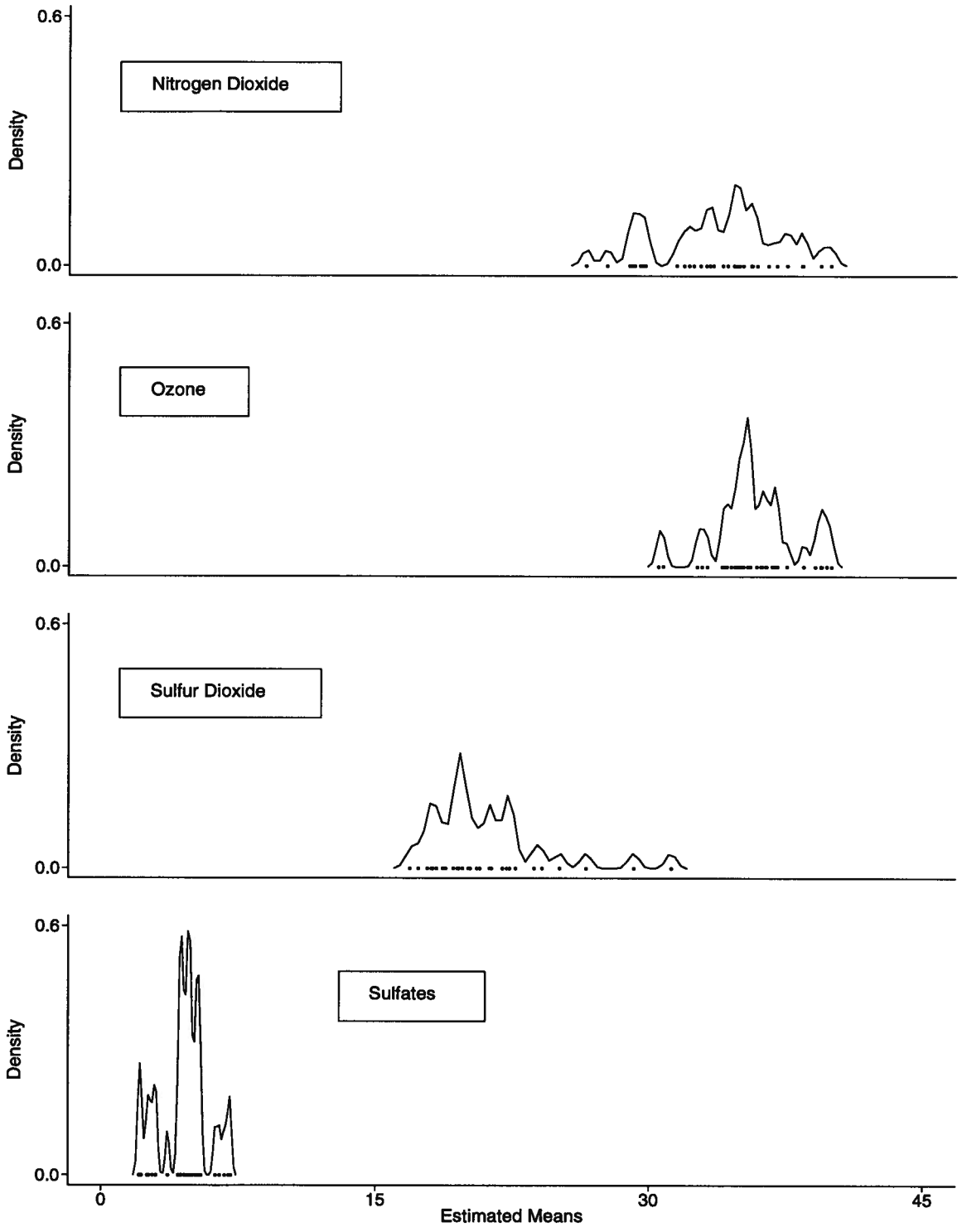

Figure 28: Distribution of the 37 estimated PHU means for the four pollutants.

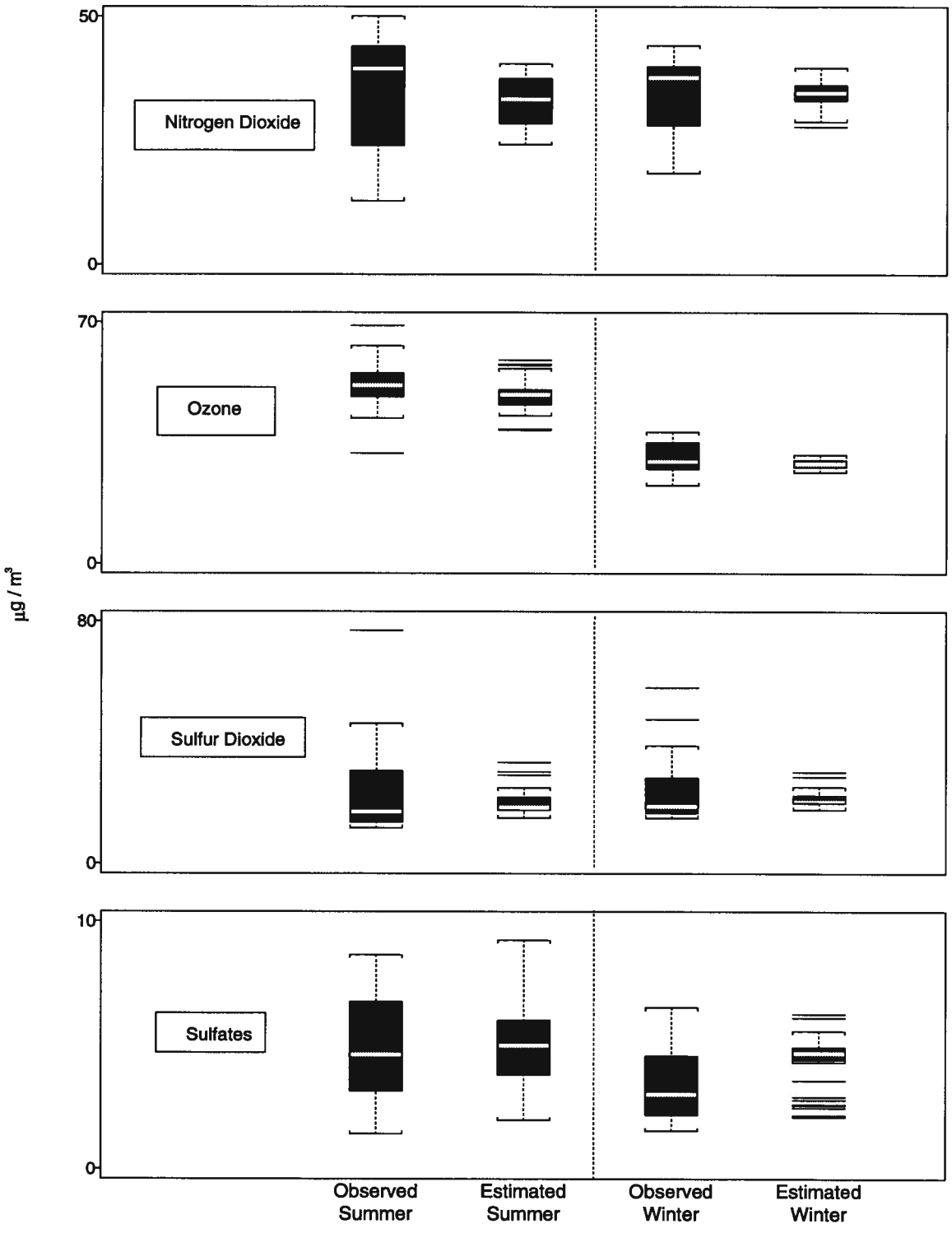

Figure 29: Comparison of pollution measurements to estimates.

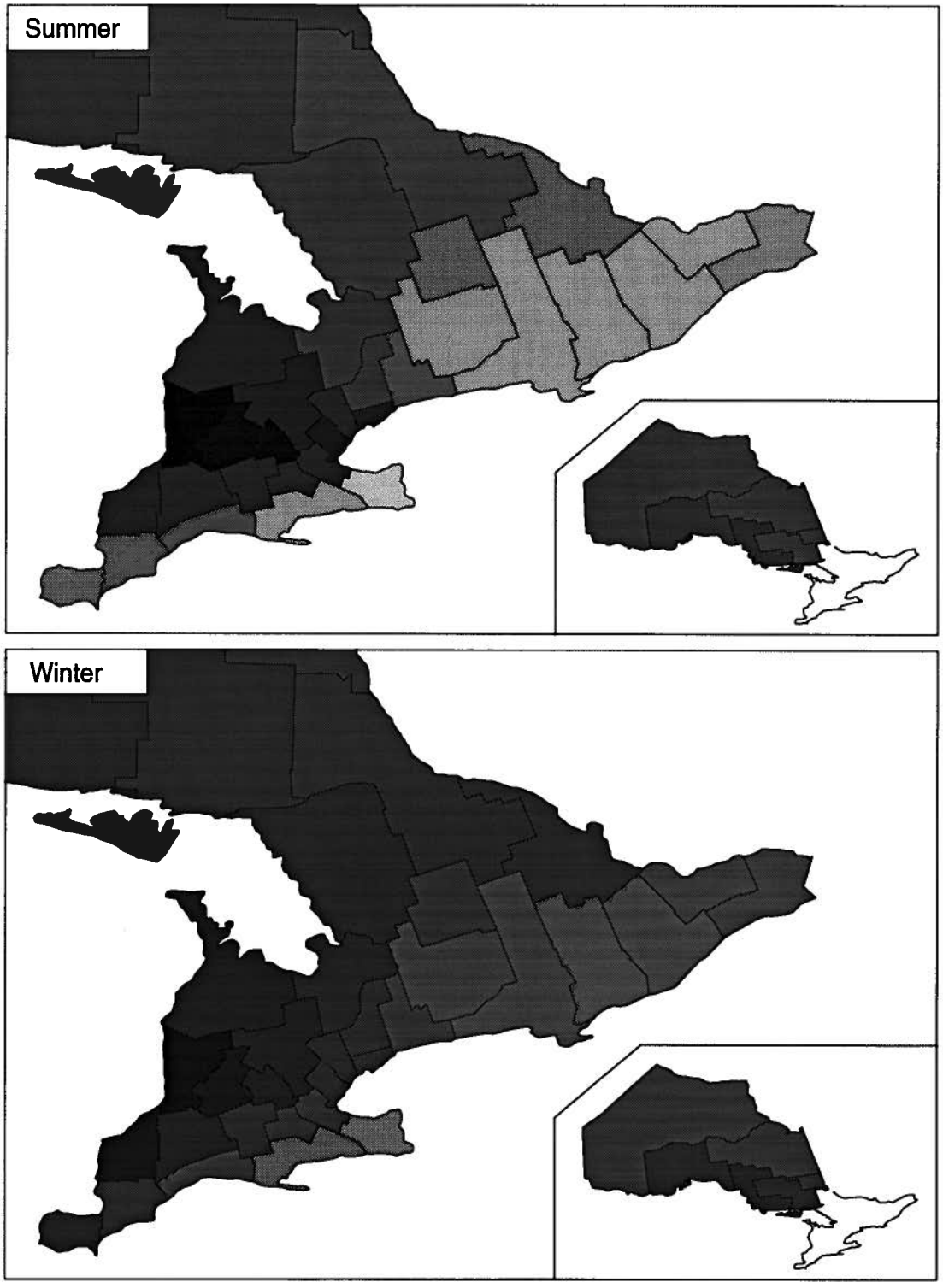

Figure 30: Estimated NO_2 six year average

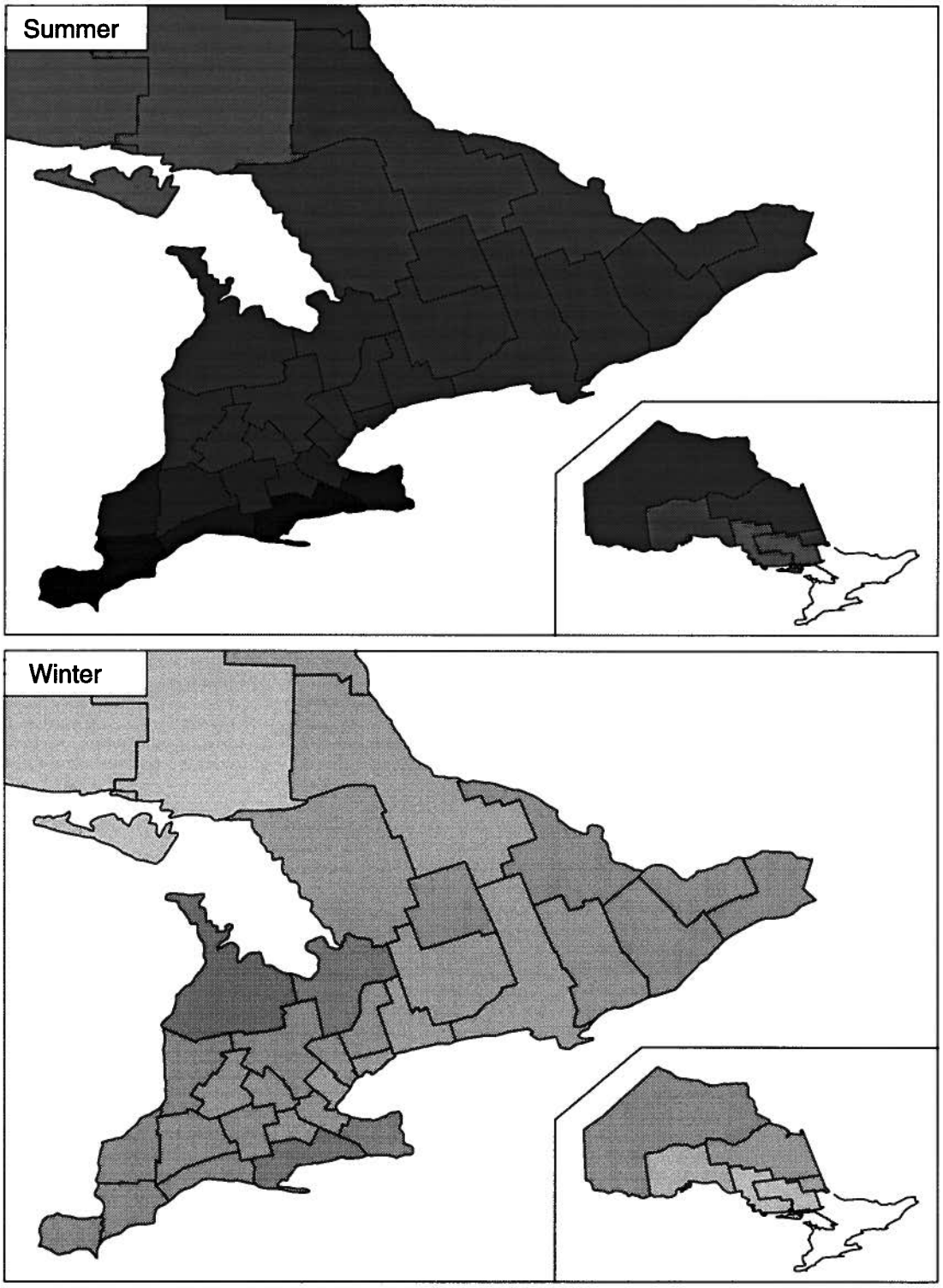

Figure 31: Estimated O_3 six year average

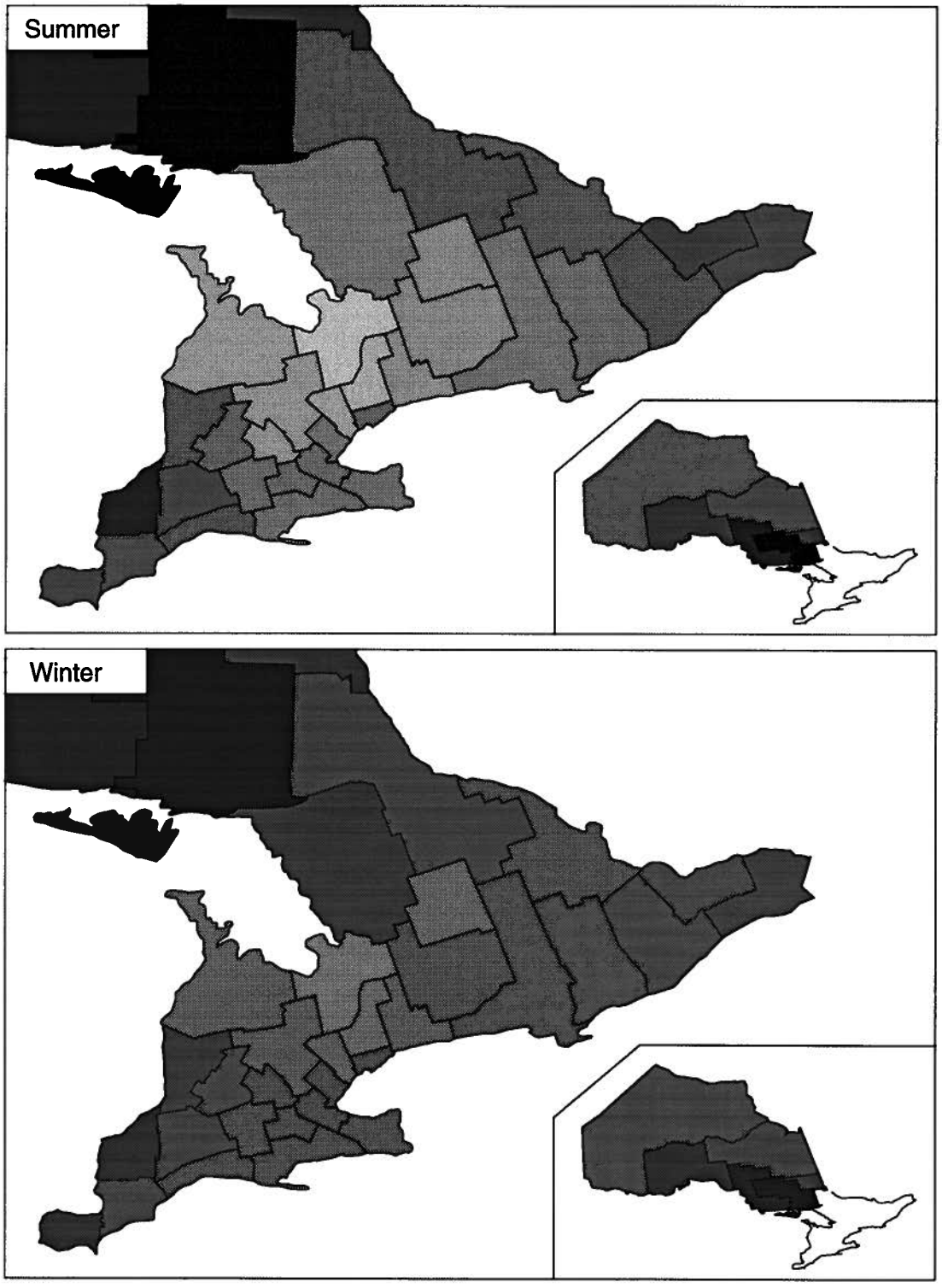

Figure 32: Estimated SO_2 six year average

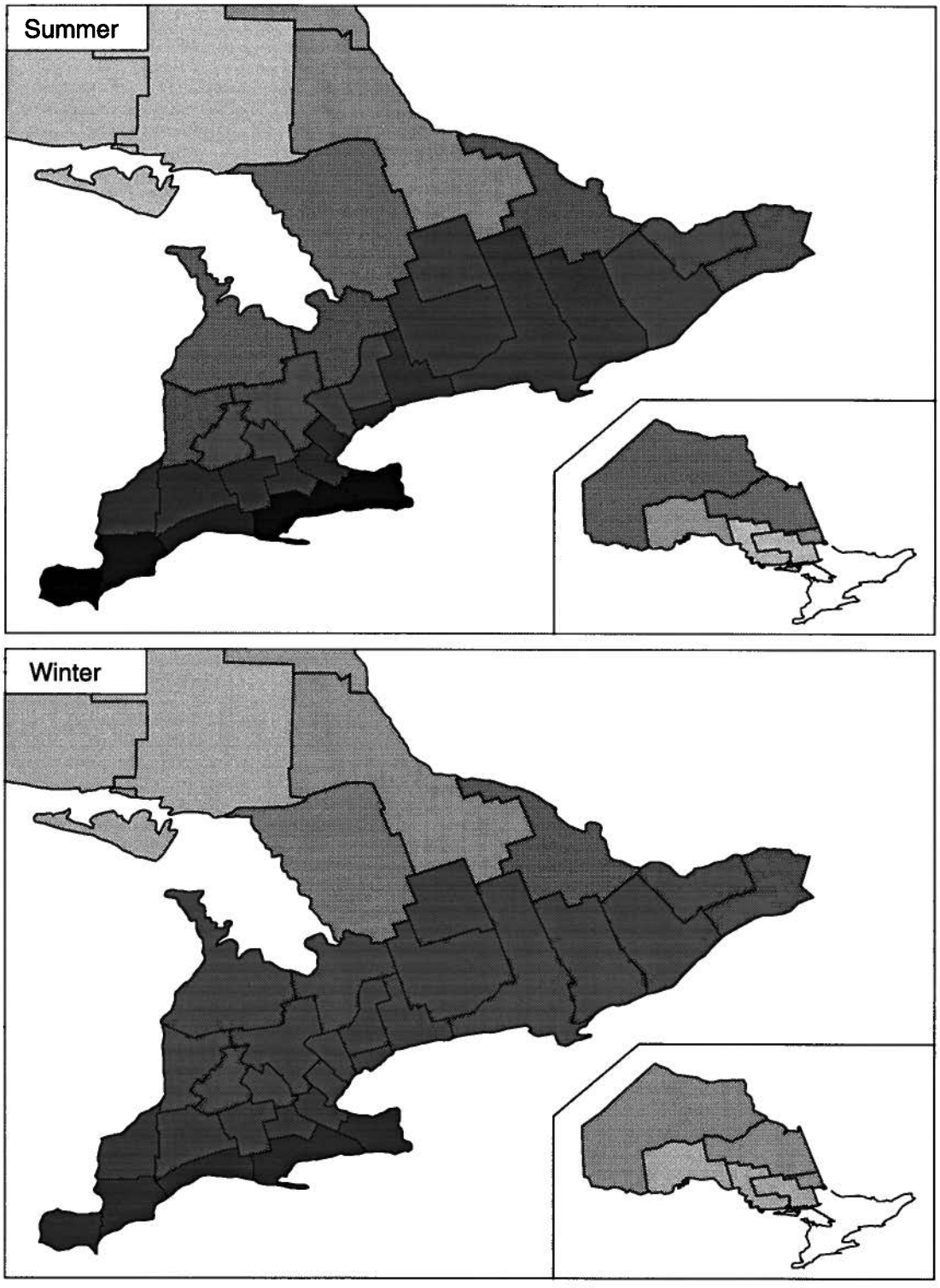

Figure 33: Estimated SO_4 six year average

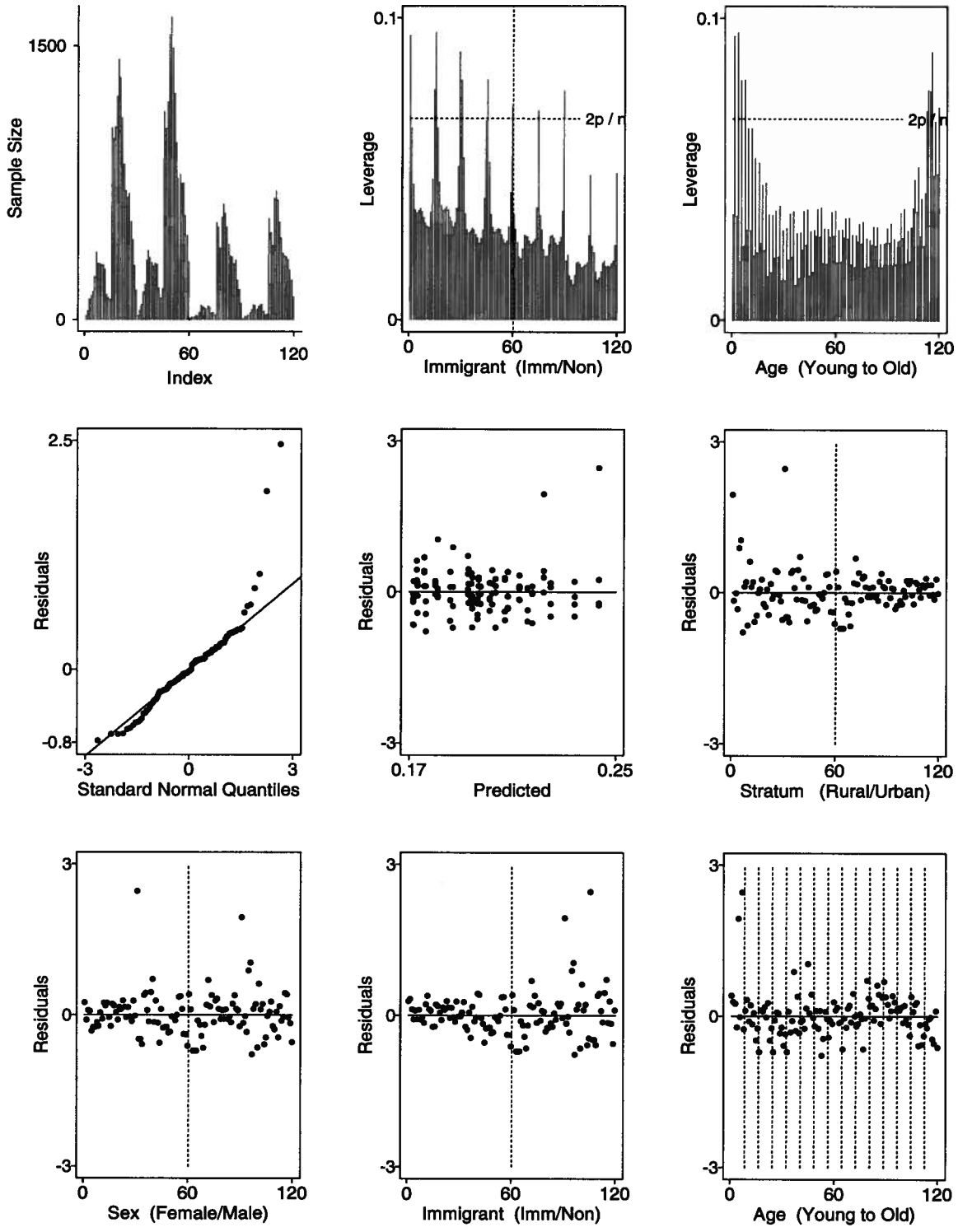

Figure 34: Asthma arcsine model diagnositcs for the demographic grouping.

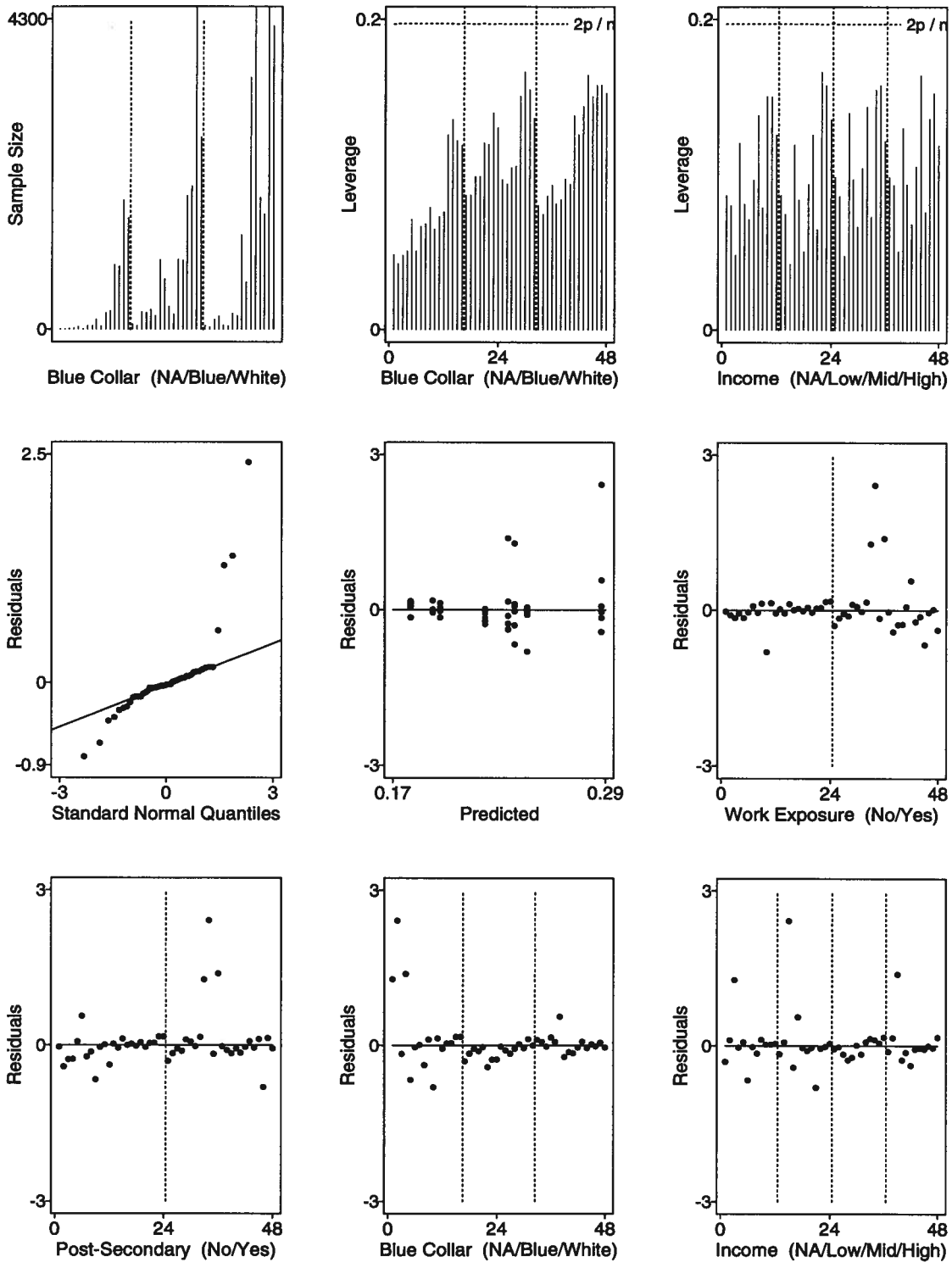

Figure 35: Asthma arcsine model diagnositcs for the socioeconomic grouping.

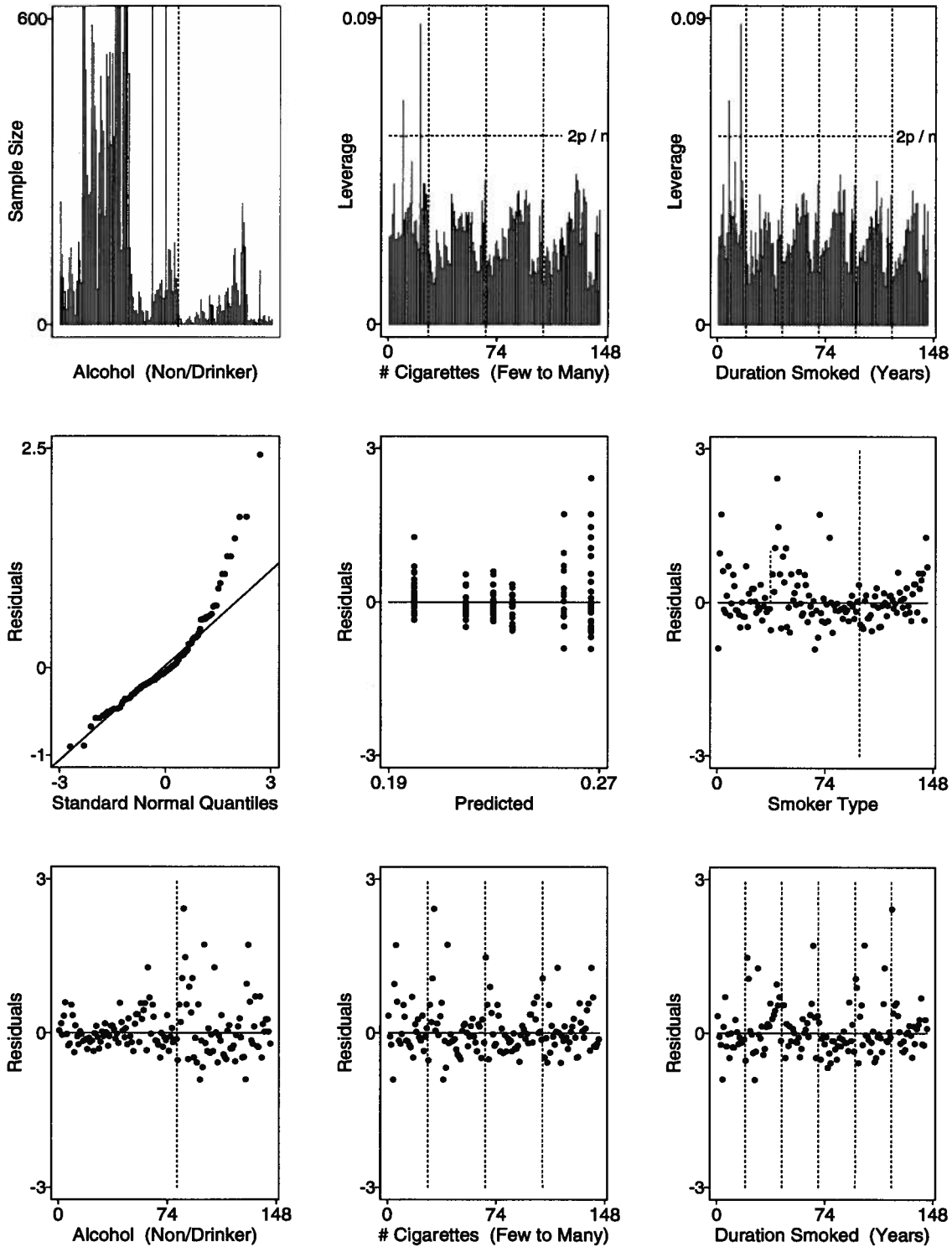

Figure 36: Asthma arcsine model diagnositcs for the lifestyle grouping.