CHAPTER 4

Setup of the Analysis and Preliminary Data Examination

4.1 Introduction

This chapter explains the organization of the analysis in the following three chapters. The first five sections give a survey of the data, and introduce the dependent and independent variables. The last two sections consist of a preliminary data analysis. In Section 4.6 the data is examined with several least squares models that are based on the assumption that the true distribution can be approximated by a continuous distribution. Section 4.7 refines the analysis by using two discrete, non-negative distributions together with a maximum likelihood method. The first model uses the Poisson distribution, but tests reveal that this distribution is inappropriate in this case because the data is overdispersed. The analysis is then repeated with the negative binomial distribution.

The program LIMDEP was the only available statistical package that supported the negative binomial distribution, and it was used for the estimation of all models in this chapter. Even though this program does not allow a very refined analysis in some cases (for example, it does not support a fixed effects model with 219 groups for the discrete distributions because of the calculation effort that this involves), it is sufficient for a preliminary analysis that emphasizes some properties of the data.

4.2 Review of the data: The Bureau of the Census data collection

The U.S. Bureau of the Census publishes monthly data on the number and the value of building permits in 21 categories of construction for municipalities in all 50 states. Data on building permits relate to the time of issuance of the permit, and not to the time of the actual start of construction. Even though construction is usually started within the same month in which the permit is issued, in some cases several months may pass before construction is started.[130] Data on building permits can therefore give only an approximation of the actual construction activity within a certain month. However, it is still the best data on construction that are available. The Census data are derived from reports by the municipalities; if reports

[130] See Census C40 Series, Appendix p. 12.

are missing (as they frequently are) then the Bureau of the Census estimates what the report would have shown, but both the actual reported numbers and the estimates are published.

As only Pennsylvania allows municipalities to tax land and buildings at different rates, this study examines only building permits that are issued by municipalities in Pennsylvania. Each municipality in Pennsylvania is classified as either a city, a borough, or a township. Each plot of land is located in only one municipality, so that it is possible to identify building activities by the name of the municipality they are located in, without running the risk of double-counting.

The Bureau of the Census data are divided into 21 categories, of which six refer to residential housekeeping construction, and the remaining 15 refer to nonhousekeeping and nonresidential construction. All categories are shown in Table 4.1.

Among the categories of nonresidential construction, the last six are most likely to be either public or tax-exempt, and can be expected to exhibt no reactions to changes in taxes; they are therefore excluded from the analysis. Categories 5, 6 and 15, which consist of additions and alterations to residential and nonresidential construction, do not involve the construction of whole units. Since such construction differs so greatly from whole units, they were examined separately from the construction of whole units.[131] In addition, close inspection of the data for these three categories reveals some eratic variations in single years, which seem to point to errors in the data set. For example, the data sets for permits for residential additions and carports in Scranton do not show a single permit in these two categories in 1990, even though in all other years several hundered permits are issued for residential additions and alterations, and between 45 and 68 permits are issued for garages and carports. It is hard to find an explanation other than an error in the data set for this and similar occurences. As the other categories do not show these obvious variations, this provides an additional reason to examine the construction of additions and alterations separately from the construction of whole units.

The determinants of residential construction differ from those of nonresidential construction, which makes it also necessary to divide the data into these two categories and to examine them separately. But there exists no apparent reason why, for example, the construction of single-family houses should respond differently to tax incentives than the construction of two-family houses, so that aggregating the first four residential and the first 9 non-residential categories into two groups is justified.

---

[131] For nonresidential construction I tested the appropriateness of dividing the data set into these categories by combining additions and alterations with the construction of whole units. The result indicated that a separate analysis is in order. See the analysis in Section 5.5.3.

Table 4.1 The construction categories in the Bureau of the Census data set

Residential Housekeeping Buildings:	(1) Single Family Houses (101)
	(2) Two-Family Buildings (102)
	(3) Three- and Four-Family Buildings (103)
	(4 Five-or-More Family Buildings (104)
	(5) Additions, Alterations and Conversions of residential buildings (105)
	(6) Additions of Garages and Carports (438)
Residential Nonhousekeeping Buildings and Nonresidential Buildings:	(7) Hotels, Motels, and Tourist Cabins (213)
	(8) Other Nonhousekeeping Shelter (214)
	(9) Amusement, Social and Recreational Buildings (318)
	(10) Industrial Buildings (320)
	(11) Parking Garages (321)
	(12) Service Stations and Repair Garages (322)
	(13) Office, Bank, and Professional Buildings (324)
	(14) Stores and Customer Service Buildings (327)
	(15) Additions, Alterations and Conversions of nonresidential buildings (437)
	(16) Churches and other Religious Buildings (319)
	(17) Hospitals and Institutional Buildings (323)
	(18) Public Works and Utilities Buildings (325)
	(19) Schools and Other Educational Buildings (326)
	(20) Other Nonresidential Buildings (328)
	(21) Structures other than Buildings (329)

Note: The numbers in brackets are the classification numbers of the Bureau of the Census. The specific contents of the data items are given in Appendix D.

The data sets have a few other mistakes, because every now and then a positive number of permits is recorded together with a value of zero. It seems to be more likely that the person who entered the data into the computer accidentally hit a number different from zero for the number of permits, than that this person entered a zero for value when the correct value was at least a four-digit number. For this reason the number of permits was set to 1 whenever the reported value was larger than zero, but it was set to 0 whenever the reported value was zero.[132]

Information about the number of permits and their value is available for all 21 categories of construction, and were obtained from the Bureau of the Census on diskettes for all years between 1980 and 1994. The data set shows for each year, for every category and for every municipality, the total number of permits that were reported throughout the year, the aggregate value of all reported permits in this year, the number of months for which reports were filed, and various identifying data including the name, county, and ID number of each municipality. For years before 1980 only the number of permits for the four categories of residential construction of whole units is available, and only on paper.[133] As Harrisburg introduced the two-rate system in 1974, it seemed appropriate to extend the data set containing the number of permits of residential construction of whole units back to 1972, to be able to examine the effect that this switch might have had on construction.

The four different groups that were examined separately from each other for different periods of time are therefore:

- Residential construction of whole units (1-4) (Years: 1972 - 1994)
- Residential additions and alterations, and garages and carports (5-6) (Years: 1980 - 1994)
- Nonresidential construction of whole units (7-14) (Years: 1980 - 1994)
- Nonresidential additions and alterations (15) (Years: 1980 1994).

[132] This data correction affected between 5 and 20 entries per data set.

[133] See Bureau of the Census, Construction Reports, "Housing Units Authorized by Building Permits and Public Contracts," (C40 series), monthly.

4.3 Municipalities used in the study

The data that were obtained on disk provided information about the building permits of 2,286 municipalities in Pennsylvania between 1980 and 1994. If all of these municipalities had reported the number of permits that were issued during all 15 years, the data set would have consisted of 32,790 entries. But it has only 23,449 entries; almost 30 percent of the observations are missing. As municipalities frequently report that they did not issue a single permit throughout a year, these missing numbers are not equal to 'zero permits', but are true missing data. The Bureau of the Census publishes its own estimates for missing information, but because the Bureau's main interest is in aggregate estimates, it is possible that their estimation method is not appropriate if one wants to examine single municipalities. In addition, the Bureau of the Census does not take any potential effects of a two-rate tax into account when missing observations for the 15 two-rate cities are estimated, so that the estimated numbers will be biased against finding a tax effect. As many observations for the two-rate cities are missing, these estimates would be of no help for this study, and they were not used in the analysis.

An additional, closely related difficulty comes from the fact that municipalities do not always report their issued permits every month.[134] The numbers that appear in the data set are the sums of permits in all reported months, but the number of months reported can be any integer from 0 to 12. Instead of restricting the analysis to municipalities that filed complete reports, every observation was included even if it was compiled from just one month, and in the analysis an adjustment was made for the missing months. A municipality was only excluded from the analysis if it did not file a report at least once during the 15 years for which data is available on disk. There were 111 places that never filed a single report, which reduced the data set to 2,175 municipalities.

A further reduction of the data set needed to be made, because 42 municipalities did not issue their own building permits throughout the whole span of time considered, but rather were part of 'Unincorporated Areas' within their respective counties for some of the time.[135] As data are not available for the whole time, these municipalities were treated as if they had always remained a part of the Unincorporated Areas. This reduced the data set to 2,133 municipalities.

---

[134] Table C.1 in Appendix C shows the number of months in each year for which data is available for the two-rate cities between 1980 and 1994.

[135] The paper edition of the Census C40 series identifies the places which belong to an Unincorporated Area in the explanatory notes to Table 4.

The next decision to be made was whether all remaining municipalities should be included in the study. Up to 1992, only 17 cities had decided to implement two-rate taxes,[136] and these cities are not a random sample of all municipalities. They differ from the majority of the municipalities in Pennsylvania in at least three ways: their population densities are very high, they lost considerable fractions of their population in the 1970s, and their per capita incomes are low. Many further important differences between Pennsylvania's municipalities can be imagined; if data were available to model all major differences, a multivariate analysis would take account of these differences. But it is very difficult to collect complete data on many variables on the municipality level, and I was only able to find information about the income, population, and the area as potential determinants of construction. Comparison of comparatively poor municipalities which are in economic distress with rich and prosperous places is therefore unlikely to show a tax effect in the former places. It seems justified to include in the analysis only municipalities that are reasonably similar to the two-rate cities.

Places with high population densities usually also have high building densities, which makes it necessary to demolish an existing structure before a new building can be erected; this increases the cost of construction, and therefore affects the decision to build. On the other hand, construction on a previously developed site can decrease the cost of connecting the site to the sewer, to electricity lines, and to water pipes. Figure 4.1 shows the distribution of population density in 1980 among the 2,116 one-rate municipalities and among the 17 two-rate cities. The population in each of the two-rate cities is between 2,000 and 8,000 persons per square mile; they are more densely populated than most municipalities. Therefore the analysis was restricted to places with a population density larger than 2,000 persons per square mile, which reduced the data set to 530 municipalities.

A second point of attention was the change in population. Figure 4.2 shows the distribution of the population change in the 1970s for the one-rate municipalities and for the two-rate cities. All of the two-rate cities lost at least 5 percent of their population, which is considerably more than most of the municipalities in Pennsylvania. Population loss can be an indicator of economic difficulties that might reduce construction activities. For that reason it seems sensible to exclude all municipalities with a change in population of more than -5 percent. In the whole data set of 2,133 municipalities, 536 places lost more than 5 percent of their population. In the reduced set of 530 municipalities, 201 had a population change in the 1970s of more than -5 percent. Thus the data set was reduced to 329 municipalities.

[136] See Figure 2.1.

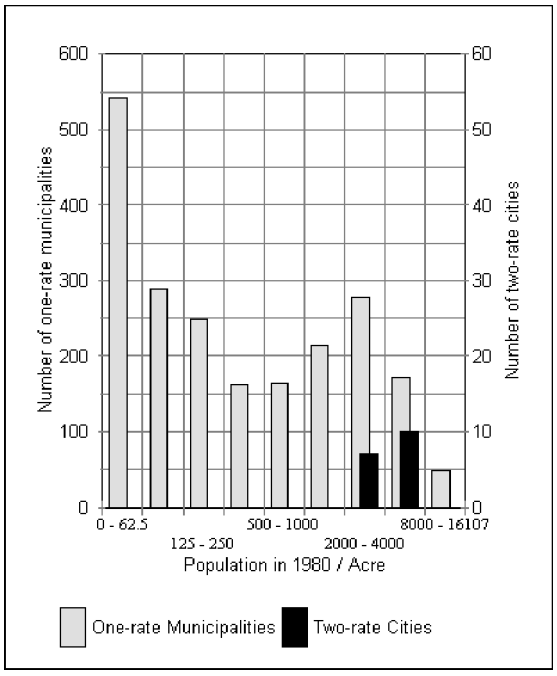

Figure 4.1 Distribution of population densities in 1980.

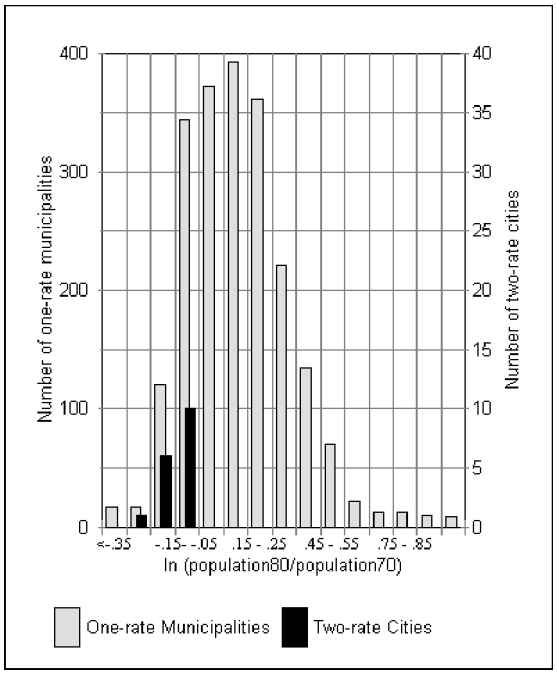

Figure 4.2 Distribution of population change during the 1970s.

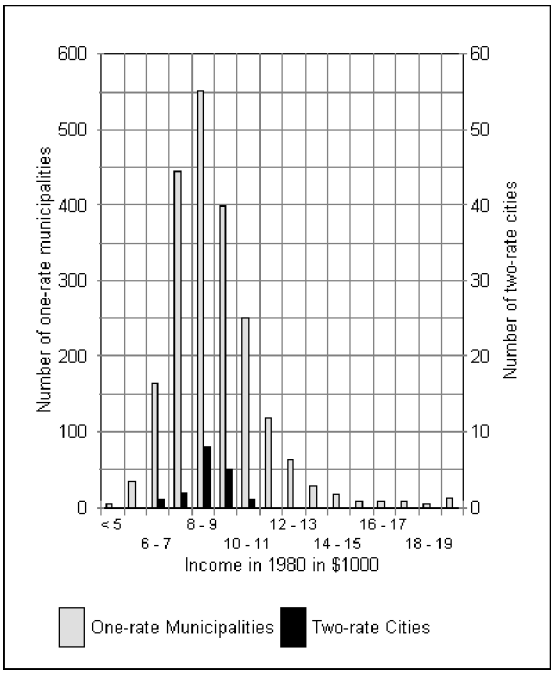

Figure 4.3 Distribution of income in 1980.

Income differences might also explain dif ferences in construction activities. First, people with lower income will tend to build fewer new houses, and second, places with low average incomes tend to have few attractive employment possibilities, so that fewer new residents are attracted. Figure 4.3 shows the distribution of 1979 per capita income for all one-rate municipalities and for the two-rate cities. None of the two-rate cities had a per capita income larger than $10,250, which made it appropriate to focus only on municipalities with a per capita income lower than that amount. But it turned out that the combination of the previous two reductions had already excluded all municipalities with a high per capita income, possibly because of the relationship between low income and population change, so that this analysis did not yield a further reduction of the data set.

A high percentage of the 2,133 municipalities had less than 2,500 inhabitants at at least one point in time during the 22 year range that was examined. Given that some of these municipalities had a population of as few as 15 inhabitants, that the smallest city with a two-rate tax (Titusville) had 6,434 inhabitants in 1990 (the year with its lowest population), and that demographic data for most places with less that 2,500 inhabitants is frequently not published, it was appropriate to restrict the analysis to places with a population larger than 2,500 people, which are 879 out of the original 2,133 municipalities. This restriction excluded an additional 110 cities from the reduced set, so that the final data set now consisted of 219 municipalities.

Uniontown and Hazleton, two of the 17 cities with a two-rate system, abolished this tax after only one or two years. It is possible that the abolition could have been predicted, so that their two-rate taxes would have created only negligible incentives for construction. For this reason these two cities were treated as if their tax ratio had been 1:1 throughout the whole time period, and only the remaining 15 cities that did not abolish the two-rate tax were considered two-rate cities.

4.4 Construction of demographic data

Data from the 1970, 1980, and 1990 Censuses were used to obtain information about the population and the income in each of the 219 municipalities in these years.[137] The three observations of income were deflated to constant (1983) dollars. Demographic information for the years between the census years is published for only a few years, and even this information comes with caveats about the unreliability of the extrapolated data, so they were not used in the analysis. It is more reliable to collect income and population data on the county level, which is available on CD-ROM for the years between 1969 and 1994 on the Regional Economic Information System (REIS) from the Bureau of Economic Analysis (BEA), and to interpolate the municipality data from the county data. As the relationship between county and municipality population and income for each of the three census years is known, the municipality data for the missing years was interpolated with the following formula

$$D_{m_t} = \begin{cases} D_{c_t} * \left(\frac{1980-t}{10} * \frac{D_{m_{1970}}}{D_{c_{1970}}} + \frac{t-1970}{10} * \frac{D_{m_{1980}}}{D_{c_{1980}}}\right) & \forall t \in [1970,1980] \\ D_{c_t} * \left(\frac{1990-t}{10} * \frac{D_{m_{1980}}}{D_{c_{1980}}} + \frac{t-1980}{10} * \frac{D_{m_{1990}}}{D_{c_{1990}}}\right) & \forall t \in [1980,1990] \\ D_{c_t} * \frac{D_{m_{1990}}}{D_{c_{1990}}} & \forall t \in [1990,1994] \end{cases} \qquad (4.1)$$

[137] See Bureau of the Census "Current Population Reports", Series P26.

where D is the demographic information of either income or population, the subscripts c and m represent county and municipality, respectively, and t gives the year to be interpolated.

Information about the area of each municipality can be found in each volume of the "*City and County Data Book*".

4.5 The dependent variables

To gain some information about the distribution of permits it is informative to look at a histogram of the observed data. Although the data in a histogram is informative only if the data are independent and identically distributed, which is probably not the case for the number of building permits, it is likely that the number of permits in similar municipalities follow similar distributions. I therefore calculated the average number of permits issued in each municipality, and used this measure to divide the municipalities into 26 groups.[138] Figure E.1 in Appendix E shows the histogram of the number of permits that were issued for each group. Within the groups the distributions of building permits for the single municipalities are similar enough to warrant the identically-distributed-assumption.

Many municipalities issue only very few permits during a year, and very often not even a single permit. For low averages the frequency of reported permits decreases with the number of permits; the distribution seems to have a long tail. For groups with a higher average number of reported permits at first the frequency increases with the number of permits, while it decreases as the number of permits becomes large. Apart from the long tail the data could be Poisson distributed, but as the histograms reveal overdispersion, an adjusted Poisson distribution or a negative binomial distribution will describe the data better.[139] The analysis in Section 4.5 is performed first with OLS, and second with a Tobit model which takes into account that the data is censored at 0. Section 4.6 examines a Poisson and a negative binomial model.

A direct comparison of the number of permits among municipalities would be meaningless, because this number naturally increases with the size of the municipality and with the number of month for which permits are reported. This makes it necessary to take account of population and months reported. Yet these should not be simply included among the explanatory variables. If two identical municipalities were combined to a single municipality, then the population of the new municipality would be the sum of the populations of the single

[138] The data set that can be obtained on disk has 23,449 entries. If the municipalities are divided in to 26 groups, each group consists of roughly 900 observations.

[139] This is supported by the results of the overdispersion tests in Section 4.7.

municipalities, while all the other independent variables (per capita income, density and population change) remained unchanged. The expected number of permits in the new municipality should be the sum of the expected number of permits of the single municipalities, which can only be achieved if the population coefficient is one and population enters the relationship multiplicatively. A similar argument can be made for the number of months examined. If for one municipality all variables stay unchanged during two consecutive years, and the only change in the analysis is that instead of 12 month now 24 months are examined, then the predicted number of permits should exactly double. 'Months reported' should therefore enter the analysis in a multiplicative relationship with the independent variables, and with a coefficient of 1 as well.[140]

If an additive relationship between the variables is assumed, and if the coefficients on population and months are multiplied by all other variables and are fixed at 1, then it is possible to divide both sides of the regression equation by population and months to create the new dependent variable 'Permits per month per person' (PMP). In the OLS regressions and in the Tobit model, PMP has been used as the dependent variable. As the number of reported permits is frequently equal to zero, it is not possible to 'log-linearize' the regression equation.

In the models which incorporate the discrete nature of the data, it is not possible to use the non-integer number PMP as the dependent variable; these models use 'Number of permits' instead. The Poisson and negative binomial models in Section 4.7 that were estimated with LIMDEP use 'Population' and 'Months reported' as independent variables without imposing the restriction that their coefficients should be equal to 1; the analysis in Chapter 5, however, imposes this restriction.

4.6 The independent variables

Because not much data is available on the municipality level, it was only possible to examine a few of the potentially explanatory factors. The main determinants of the decision to request a building permit which can be calculated with the available data are

- income relative to average income
- population density
- population change in the previous period(s)
- adjusted tax differential
- demolition of buildings in the previous period(s)

[140] This relies on the assumption that the failure to report all data is randomly distributed, and that the missing months do not typically describe months with construction activity that is above or below the yearly average. Yet there is no indication that the assumption of randomness is not valid.

Instead of considering income relative to average income, it would be possible to consider absolute income. Yet the use of absolute income might lead to an under-estimation of the influence of income if average real income rises over time; by using relative income this potential bias can be avoided.

The influence of the two-rate tax is measured as the adjusted tax differential. The tax differential is given by the difference between the city tax on land and the city tax on buildings, which is 0 for all municipalities with a one-rate tax, and positive for the two-rate cities. But cities *assess* property at different fractions of its market value; among two-rate cities land is assessed at rates between 6.25 percent (Coatesville) and 53.66 percent (Scraton) of the market value of buildings. This makes it necessary to multiply the tax differential by the assessment ratio to compute the adjusted tax differential as a percentage of the market value. The adjusted tax differential varies between 0.16 percent in Coatesville since 1991, and 12.17 percent in Aliquippa in 1993. The ratio of the tax rate on land to the tax rate on structures in alternative measure of the tax differences, which has been used by Mathis and Zech (1982). But a correct tax ratio would have to take the school district tax and the county tax into account, because all counties and almost all school districts tax structures at the same rate as land,[141] which would make the true tax ratio smaller than it appears if one just considers the city tax. In addition, while one can easily determine the county to which a municipality belongs, it would be very tedious to determine the actual borders of the school districts. As data on school district and county tax rates were not readily available, I decided to investigate only the city tax, and to use the adjusted tax differential as the tax variable.

The direction of causation for the first four determinants ought to be unambiguous; it is implausible that a higher number of permits issued will cause the relative income to increase immediately, or that today's building activities have caused population change in the past. But this relationship is not so obvious for the demolition of buildings. If many buildings have been demolished in the past, then more empty lots and fewer structures exist, which should lead to higher building activity. On the other hand, the higher the building activity, the more empty lots are needed for construction, and the more structures need to be demolished. For this reason it seemed to be inappropriate to use the demolition of structures as an independent variable. A second argument against its use stems from the possibility that demolition might be highly correlated with the impact of population density on building permits. The higher the density, the more existing structures usually have to be torn down in order to build a new building, and the more expensive construction becomes. As population density proved to be highly significant, it was necessary to exclude demolition from the analysis.

[141] Since 1993 the school district of Aliquippa has taxed land and structures at different rates.

4.7 Data analysis with OLS and the Tobit model

As preliminary analyses, I examined several models that assume a continuous distribution of the error term. As these models do not take account of the discrete nature of the data, they are used merely as an introduction, and only the results for residential construction of whole units are shown.

The first column in Table 4.2 shows the estimated coefficients for the OLS model. The estimated equation is

$$PMP_{i,t} = \beta_1 + \beta_2 \cdot D_{i,t} + \beta_3 \cdot I_{i,t} + \beta_4 \cdot \dot{P}_{i,t} + \beta_5 \cdot YD_t + \beta_6 \cdot ATD_{i,t} + \epsilon_{i,t} \; . \tag{4.2}$$

where $D_{i,t}$ is density, $I_{i,t}$ is income relative to average income, $P_{i,t}$ is population change, the adjusted tax differential is given by $ATD_{i,t}$ and YD_t is the yearly dummy.[142] All parameters except the adjusted tax differential are statistically significant. The Durbin-Watson test

Table 4.2 Estimation results of models that assume a continuous distribution

	OLS	OLS (Hetero-skedastic)	Fixed Effect	Tobit	Tobit (Heteroskedastic)
Constant	0.29773 (0.0202)	(0.0269)	-	0.34984 (0.0246)	0.27460 (0.0407)
ATD	-0.003166 (0.01196)	(0.0039)	0.00851 (0.0148)	-0.00169 (0.0158)	-0.00391 (0.00576)
Density	-0.10022 (0.01412)	(0.0125)	-0.19890 (0.1445)	-0.20830 (0.0182)	-0.14349 (0.0170)
Income	0.076036 (0.01764)	(0.0148)	(0.11362) (0.0680)	0.14856 (0.0216)	0.10380 (0.0186)
PopChange	2.8627 (0.4769)	(0.0538)	1.2324 (0.5665)	4.2736 (0.5805)	2.1085 (0.2787)
R^2 or Loglikelihood	0.06255	-	0.1337	-1352.736	-847.2008

Note: Standard errors are shown in parentheses. Coefficients of yearly dummies are not shown.

[142] As the coefficients of the yearly dummies are only of minor interest, they are not shown for any regression.

statisic for autocorrelation is 1.91, which indicates that the error terms are relatively independent of each other. As the histograms in Figure E.1 in Appendix E reveal that the data is heteroskedastic, the second model uses White's estimator for the covariance matrix to estimate efficient coefficients.[143] As before, all coefficients except the tax parameter are significant. But the R^2 indicates that the explanatory power of the model is rather poor. The third model in the third column of Table 4.2 therefore uses a dummy variable for each municipality (fixed effects model); the R^2 increases slightly, and none of the variables is significant anymore. But both of these models fail to take into account that it is impossible to issue a negative number of permits, which leads to a sample selection bias. The Tobit (or 'censored regression') model avoids this sample selection bias. The general formulation of the Tobit model is[144]

$$\begin{aligned} y_i^* &= \beta' x_i + \epsilon_i \ , \\ y_i &= 0 \qquad if\ y_i^* \le 0 \ , \\ y_i &= y_i^* \qquad if\ y_i^* > 0 \ , \end{aligned} \tag{4.3}$$

so that the error term is assumed to have a censored normal distribution.

Unfortunately LIMDEP does not permit the estimation of a Tobit model as a fixed effects model with 219 groups, so the results in the fourth column in Table 4.2 were obtained without dummies for municipalities. Again, every coefficient except the coefficient on taxes is significant, and the signs of the coefficients are the same as in the OLS regression. Correcting the covariance matrix for heteroskedasticity increases the standard deviations in column five, but none of the coefficients changes its significance, while the loglikelihood improved by 505.54.[145]

Various authors have shown that the analysis of count data is improved by the use of discrete distributions, for example the Poisson and the negative binomial distribution, which are used in the analysis in the next section.

---

[143] See White (1978), and Green (1990), p. 391.

[144] Green (1990), pp. 694 - 697.

[145] The results in this section also apply to the other three data sets (residential additions and alterations, nonresidential whole units, and nonresidential additions and alterations). But during the analysis it became apparent how easy it is to 'produce' a significant coefficient for the tax differential, and how crucial the functional form proves to be: if the regressions are repeated with 'permits' instead of 'PMP' as the dependent variable, and 'Population' and 'Months reported' are included as additive and unconstrained independent variables, most data sets show a positive and significant tax coefficient for the OLS and the Tobit model; in the data set of residential construction of whole units the R^2 in the OLS model increases to 0.228 and in the fixed effects model even to 0.668.

4.8 Data analysis with the Poisson and the negative binomial model

In the Poisson model the single parameter λ_i is equal to the expected value of the Poisson distribution, and the independent variables are introduced into the model by expressing λ_i as a deterministic function of these variables.[146] In order to guarantee a positive expected value, the functional form estimated by LIMDEP (and in the literature) is $\lambda_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i)$, where $\boldsymbol{\beta}'$ is the parameter vector and $\mathbf{x}_i$ is the vector of independent variables. As 'Population' and 'Months reported' should enter the equation multiplicatively, their logarithms are used, which yields the following equation:

$$\begin{aligned}\lambda_{i,t} &= e^{\beta_1 + \beta_2 \ln(pop_{i,t}) + \beta_3 \ln(m_{i,t}) + \beta_4 D_{i,t} + \beta_5 I_{i,t} + \beta_6 \dot{P}_{i,t} + \beta_7 YD_t + \beta_8 ATD_{i,t}} \\ &= (pop_{i,t})^{\beta_2} \cdot (m_{i,t})^{\beta_3} \cdot e^{\beta_1 + \beta_4 D_{i,t} + \beta_5 I_{i,t} + \beta_6 \dot{P}_{i,t} + \beta_7 YD_t + \beta_8 ATD_{i,t}}\end{aligned} \tag{4.4}$$

Table 4.3 Estimation results of the Poisson model

	Residential Whole Unts	Residential Additions and Alterations	Nonresidential Whole Units	Nonresidential Additions and Alterations
Constant	-6.8557 (0.0696)	-3.8140 (0.0330)	-9.4765 (0.2130)	-1.3319 (0.0810)
ATD	-0.04713 (0.0062)	0.22377 (0.0188)	-0.0999 (0.0208)	0.19524 (0.0385)
Density	-0.5437 (0.0135)	-0.51052 (0.0081)	-0.99142 (0.0615)	-4.9922 (0.1774)
Income	0.4597 (0.0128)	0.58282 (0.0066)	-0.50331 (0.0541)	0.26379 (0.0138)
PopChange	15.828 (0.2811)	15.284 (0.2681)	3.6150 (1.935)	32.813 (0.5932)
LogPop	1.0403 (0.0023)	0.77009 (0.0013)	0.89598 (0.0104)	1.0548 (0.0029)
LogMonth	0.556 (0.02733)	0.81976 (0.01268)	0.78024 (0.0710)	1.6931 (0.0316)
Loglikelihood	-62,652.94	-17,4457.90	-5,430.49	-44,519.32

Note: Standard errors are shown in parentheses. Coefficients of yearly dummies are not shown

[146] See Section 3.2.

It should be expected that the estimated coefficients of 'Population' and 'Months reported' will both be close to 1.

Table 4.3 shows the results for all four data sets. All coefficients are highly significant; the tax coefficient is positive for both data sets for additions and alterations, and negative for the construction of all whole units. Note that the coefficients of population size is almost equal to one, as predicted. The coefficient of the number of months reported is at least in the vicinity of 1.

The use of the Poisson model is only appropriate if the data have null dispersion, that is, if the mean of the dependent variable is equal to its variance. Cameron and Trivedi (1990) developed the following overdispersion test, which tests the null hypothesis H_0: $Var(y_i) = E(y_i)$ versus the alternative hypothesis H_A: $Var(y_i) = E(y_i) + \alpha f(E(y_i))$. They propose two different functional forms for $f(E(y_i))$: the first function is the 45-degree line with $f_I(E(y_i)) = E(y_i)$, and the second function is quadratic in the mean with $f_{II}(E(y_i)) = E(y_i)^2$. Table 4.4 shows the estimated values for α for both functional forms: f_I is supported by all four data sets with a positive and significant value of α, while f_{II} is still signifi cant for all four data sets, although the coefficient for nonresidential construction of whole units is only barely significant. The histograms in Figure E.1 already indicated that the data are overdispersed; these tests now confirm that the Poisson model is not correct, and that the negative binomial model should be used instead.

One possibility for deriving the negative binomial distribution is to use a Poisson distribution whose parameter λ_i follows a gamma distribution.[147] This gamma distribution can be parameterized so that one of the two parameters of the negative binomial distribution is its expected value μ_i. To ensure non-negativity of μ_i, the functional form estimated by LIMDEP

Table 4.4 Estimation results of overdispersion tests

	Residential Whole Units	Residential Additions and Alterations	Nonresidential Whole Units	Nonresidential Additions and Alterations
f_I: α	45.875 (3.925)	178.45 (25.8137)	2.9659 (0.6443)	45.451 (7.4397)
f_{II}: α	0.13152 (0.0268)	0.50252 (0.0860)	0.28348 (0.1102)	0.25702 (0.0622)

Note: Standard errors are shown in parentheses.

[147] Compare Section 3.4.

Table 4.5 Estimation results of the negative binomial model

	Residential Whole Units	Residential Additions and Alterations	Nonresidential Whole Units	Nonresidential Additions and Alterations
Constant	-6.9553 (0.3042)	-6.8990 (0.3138)	-8.9394 (0.3847)	-0.4762 (0.2087)
ATD	-0.12153 (0.0689)	-0.05829 (0.0928)	-0.002049 (0.0695)	-0.09644 (0.1227)
Density	-1.1311 (0.0700)	0.24681 (0.0833)	-1.0107 (0.0920)	-2.9646 (1.154)
Income	0.79347 (0.1040)	0.29188 (0.0979)	-0.19583 (0.0970)	0.33915 (0.1180)
PopChange	24.578 (1.8170)	0.92304 (3.729)	3.7360 (3.1390)	16.832 (3.107)
LogPop	1.0314 (0.0255)	0.93448 (0.0262)	0.84105 (0.0265)	1.1374 (0.0338)
LogMonth	0.66102 (0.0615)	1.1420 (0.0524)	0.81909 (0.0980)	1.2458 (0.0770)
LogLikelihood	-14,272.93	-13,881.17	-4,467.00	-8,933.61

Note: Standard errors are in parentheses. Coefficients of yearly dummies are not shown.

is $\mu_i = \exp(\boldsymbol{\beta}'\mathbf{x}_i)$, and the estimated equation is again given by equation 4.2. Table 4.5 shows the estimated coefficients for all four data sets. Contrary to the results of the Poisson model, the adjusted tax differential does not have any significant impact in any data set. In addition, while in the Poisson models all coefficients but the coefficient for population change with nonresidential construction of whole units were significant, now several coefficients in every data set are not significant anymore.

4.9 Conclusion

The analyses in this chapter yielded different results that strongly depend on the assumptions of the different models. The probably most appropriate model for this data uses the negative binomial distribution, and the estimation did not show any significant impact of the two-rate tax.

Yet the current analysis has several potential methodological problems: The second parameter of the negative binomial distribution is usually used to describe a relationship between the mean and the variance of y_i. Even though a variety of relationships is possible, LIMDEP uses the functional form VAR (y_i) = E (y_i) + α E $(y_i)^2$, which was identified by Cameron and Trivedi (1986) as appropriate for *their* data, and has been used ever since, in all the published applications that are cited in Chapter 3, without exception. But the overdispersion test did not indicate that this formulation is better than the other formulation for this data, so that other relationships (which in effect would describe a different model) might yield different results.

The OLS results already indicated that it might be more appropriate to use a fixed effects model, but LIMDEP does not support these models together with the negative binomial distribution. Although the OLS model did not show many signs of autocorrelation, it used a different dependent variable than the two discrete models. It is possible that the division of the number of permits by the population size and the number of month reported removed existing serial correlation, which needs to be taken into account in the discrete models. The analysis in Section 5.3 uses a fixed effects model, and examines the impact of serial correlation.

It is also worth asking, whether the negative binomial distribution is the best approximation of the true distribution of the data. Its use is generally justified by two facts: first, if the data show signs of overdispersion, because the negative binomial distribution is able to deal with overdispersion in a straightforward fashion. Second, the negative binomial distribution is the only distribution that is available in closed form solution as a mixed Poisson distribution. Mixing the Poisson distribution with any other distribution but the gamma distribution makes a maximum likelihood estimation extremely tedious.[148] Fortunately, another estimation procedure proves to be less cumbersome. In Chapter 6 the data are examined with a Markov Chain Monte Carlo method, which makes it possible to use a model that combines the Poisson with different distributions; in Chapter 6 a lognormal distribution will be used instead of the gamma distribution.

---

[148] For several weeks I experimented with a generalized Poisson distribution, for which no closed form solution is available, and whose maximum likelihood estimation was possible only under great difficulties. A brief description of this approach is given in Appendix F.