Estimating Latent Effects in Commercial Property Models

BY CARMELA QUINTOS, PH.D., MAI

New York City annually values commercial properties by the income approach. Commercial properties with an assessed value greater than $40,000 are required to file income and expense statements with the Department of Finance. Some of these required filers may apply for exclusion from filing or they may choose not to file and instead pay a penalty. There are also voluntary filers, who are not required to file but nevertheless submit statements. The filings received are used to formulate income and expense regression models. These models are used to develop comparable rental models and to formulate assessment guidelines based on location and physical characteristics.

Regression models that are used for assessment represent relationships between a response variable, in our case income or expense, and property characteristics and are categorized by the International Association of Assessing Officers (IAAO) as additive (linear), multiplicative (linear in logs) or hybrid (nonlinear) (IAAO 2003, 7). In an additive specification, the contributory effects to income or expense are added together. In a multiplicative model, income or expense is affected by a base rate multiplied by contributory property-related factors. The hybrid model accommodates both additive and multiplicative effects. In all three specifications, the independent variables are observed factors, such as physical characteristics, or measured variables, such as neighborhood ratings. Any unobserved and unmeasured effects are captured in the residuals.

If there are unobserved and unmeasured *underlying* effects that affect the response variable, then these variables are *latent*. Latent variables do not lend themselves to representation in a measured variable, such as class rating or square footage, but are instead *underly-*

Carmela Quintos, Ph.D., MAI, *is Assistant Commissioner of Property Valuation and Mapping, Department of Finance, City of New York. She received her doctorate in economics from Yale University, New Haven, CT, and has held faculty positions at the University of Rochester (NY) and Washington University in St. Louis. Her research interests center on applications of spatial and time series techniques to asset valuation. She has practiced valuation of residential and commercial properties and is a New York State Certified Assessor. The views expressed in this article are those of the author and do not reflect the views of the City of New York or the NYC Department of Finance.*

ing factors, such as a common trend, unspecified correlation, or unmeasured heterogeneity. Latent variables are random variables whose realized values are hidden. Their properties must thus be inferred indirectly using a statistical model connecting the latent (unobserved) variables to the observed variables. If these latent effects are not accounted for in the model framework, they are pushed to the residuals and the prediction errors increase. Latent variables occur not because independent variables are omitted but because there is nothing in the model framework that specifies how to "discover" the underlying factor from observables.

Consider a hybrid model of income and expense. A hybrid model is a nonlinear model of income (or expense) regressed on multiplicative and additive factors. Examples of multiplicative factors used in this article are location and class rating; examples of additive factors used in our specification are building and land square footage, age, and stories. An important factor is lease information, which we know affects income, but we have no data. Even though the Department of Finance collects income and expense filings, the details of the lease—such as the term, rent escalation, expense stops, and the like—are not reported. Because leases are *property specific* but are unknown, they are pushed to the residual as an *unobserved property-specific effect* (unobserved firm heterogeneity). Thus we know, and will show later, that the hybrid model framework is misspecified since there is nothing in the specification about estimating unobserved property-specific effects. Instead, we work with a hybrid model in a panel data framework which allows estimation of unobserved property-specific effects without specifying their cause.

Panel data models can be used to describe the behavior of firms (or properties) over time. They are widely applied in the social sciences for their capacity to capture the complex relationships that occur within a group across time, and between groups across time. (See Baltagi [2008] and Hsiao [2003] for a review.) Having both the cross-section and time dimension allows us to specify group effects, or heterogeneity between groups, in different forms. This article demonstrates a latent variable model called a random effects model (also called a mixed, multilevel, hierarchical model) to estimate the unobserved heterogeneity (the latent effect). A random effect is essentially a group effect that can be measured by a deterministic part (the mean of the group) and a stochastic part that affects only that group. The random effect is the latent variable and the spreads in intercepts or slopes capture the heterogeneity in the population across specified groups.

In this article, we apply the random effects model to the income and expense filings of office buildings in Manhattan for years 2010–2013. We test different specifications of group effects—between firms in the multiplicative part of the hybrid specification (scale effect) and between office classes, between office areas, and between firms in the additive part of their specification (level effect). The results show that allowing the intercept to vary by office property in the additive part of the hybrid specification improves income and expense estimates.

We also discuss a method to extend latent effect estimation to nonfilers. Since the estimated latent variable, i.e., the spreads in intercepts for office properties around an overall mean, is estimated for each filer, we can map the latent effect on a surface by location (i.e., by its *x-y* coordinate). By interpolating the surface using spline, we get a continuous surface. Thus for any office property in Manhattan, we can take its *x-y* coordinate and find its latent effect estimate from the surface. This interpolated latent effect can be entered into the income and expense hybrid equation, together with

physical characteristics, to get an income and expense estimate for nonfilers.

Random effects models have not been widely used in the field of mass appraisal. A hierarchical model application to the estimation of property prices in the Netherlands is discussed in Francke and Vos (2004) and Francke (2008). Manhattan commercial models of sale prices using IAAO's hybrid specification are explored in O'Connor, Panos, and Som (1991). For models of income and expense, however, we are not aware of a model in a random effects (panel data) framework—most likely due to the lack of property-level data of income and expense filings. Thus this article not only contributes to the literature in the use of latent variable (panel data) models to estimate income and expense, but also extends it by illustrating an out-of-sample procedure, based on interpolation methods, to predict latent effects.

The rest of this article describes our methods and results and is organized as follows. The next section sets up our model framework—the hybrid-panel data model and its random effects specification. The following section uses the hybrid-panel data framework to estimate income. Various random effect models are tested to conclude to a specification that best predicts income. The subsequent section goes through similar steps but with expense ratios as the response variable. A model for expenses incorporating time-varying effects is discussed in this section. The following section discusses the extension to nonfilers. It describes surface interpolation of the latent effect variable so that the surface can be used for any property in Manhattan with an *x-y* coordinate (whether filer or nonfiler). The final section summarizes the results and contribution of this modeling method.

Model Framework

Market analysis using the IAAO hybrid model structure is most efficiently supported by nonlinear regression. The hybrid formula is designed to take into account effects that enter as additive, which we denote by *Z*, and multiplicative, which we denote by *X*. It is typically applied to the estimation of market value using the basic formula:

(1)

$$MV = \Pi X_q \times [(\Pi X_B \times \Sigma Z_B) + (\Pi X_L \times \Sigma Z_L) + \Sigma Z_O]$$

where

MV is the estimate of market value,

ΠX_q is the product of general qualitative variables,

ΠX_B is the product of building qualitative variables,

ΣZ_B is the sum of building additive variables,

ΠX_L is the product of land qualitative variables,

ΣZ_L is the sum of land additive variables,

ΣZ_O is the sum of other additive variables (IAAO 1990, 351; Gloudemans 1999, 121).

The nonlinear framework permits multiplicative adjustments only to the land (topography) or building (size adjustment) or the overall property (location).

When market value is derived by the income approach, the formula is

(2)

$$MV = \frac{NOI}{CapRate} = \frac{Income - Expense}{CapRate}$$

where *NOI* is net operating income and *CapRate* is the capitalization rate. Thus *MV* is derived from regression models of income, expense, and capitalization rates. Modeling income and expense from a population of filers is the focus of this article, while modeling of capitalization rates, which are determined from a separate population of filers and sold properties, is not covered here.

Income and Expense Model using Single-year Data

Each year firms or owners are required to file income and expense statements on their properties. Regressions based on the filings are used to formulate income and expense models. These models are used to value nonfilers, which are modeled as out-of-sample predictions, and to develop guidelines based on location and physical characteristics.

Our data consist of office filers for Manhattan. The model framework is

(3)

$$Y = \Pi X_q \times (\Sigma Z_B + \Sigma Z_L) = \Pi X \times \Sigma Z$$

where Y is either income per square foot (psf) or the expense ratio which is defined as *ExpRatio = Expense/Income.* Note that our model, equation 3, is a special case of equation 1 in which we include only overall property multipliers such as location. Taking logarithms, the estimating equation is

(4)

$$\ln(Y) = \beta_0 + \beta_1 \ln(X_1) + \cdots + \beta_M \ln(X_M) + \ln(\alpha_1 Z_1 + \cdots + \alpha_A Z_A) + e$$

where there are M qualitative adjustments to the total A building and land additive variables.

For example, to measure income, the left-hand side would be $\ln(\textit{income psf})$ regressed on the qualitative adjustments—office area and office class; the building additive variables—age, office square footage, retail square footage, residential square footage, and number of stories; and the land additive variable—land size.

Panel Data Models to Capture Latent Effects (Unobserved Heterogeneity)

The regression model consists of the observed or measured variables, X and Z, and the residuals, e, which contain the unobserved factors. It is assumed that all variables that are measured and significant are included in the specification so the residuals contain the best fit given known regressors X and Z. The goal is to measure unobserved factors that are known to affect income or expense (and therefore should not be in the residual) but do not lend themselves to representation by a measured variable such as the independent variables X and Z. Such variables are called latent variables. Examples of latent variable models are random coefficient models allowing β and/or α to vary by, say, time or location. The random effect is the latent variable and the spreads in intercepts or slopes capture the heterogeneity in the population across time, space (location), between firms (office properties), or within any other grouping. Estimating latent variables is, in effect, "discovering" underlying relationships or factors (heterogeneity) that would otherwise be in the residuals (unobserved and unmeasured). The use of latent variable models acknowledges measurement problems in defining a complete set of observable regressors X and Z.

To capture heterogeneity across firms, we specify a panel data model. Panel data comprise a dataset that combines time series and cross section. For N office properties and T filing periods, the dataset has sample size $N \times T$. For each firm i that files in period t, the regression model is

(5)

$$\ln(y_{it}) = b_0 + b_1 \ln(x_{1it}) + \cdots + b_M \ln(x_{Mit}) + \ln(a_1 z_{1it} + \cdots + a_A z_{Ait}) + e_{it}$$

where the error e_{it} is assumed to be Normal and independent and identically distributed (iid). When the iid or Normal assumption fails, then latent variable models can capture the unobserved effects in the residual. We will consider models in two categories: pooled regression and random effects regression.

Pooled Regression

A nonlinear regression is run on equation 5 using the full sample $N \times T$. In this case, the coefficients b and a are the same across time and across office

properties so there is no heterogeneity that is estimated.

Random Effects Regression

A random effect is a group effect that can be measured by a deterministic part (the mean of the group) and a stochastic part that affects only that group and can be independent or correlated with the overall error e_{it}. Random effect models are a class of latent variable models designed to capture heterogeneity across groups (the latent effect). We will consider four random effects models in the following sections.

Random Effect Across Office Properties (i) (or Firm Group Effect Across i)

The model measures heterogeneity across office properties in the multiplicative part of the specification. It is specified as

(6)

$$\ln(y_{it}) = b_{0i} + b_1 \ln(x_{1it}) + \cdots + b_M \ln(x_{Mit}) + \ln(a_1 z_{1it} + \cdots + a_A z_{Ait}) + e_{it}$$

$$b_{0i} = c_0 + u_i$$

where the error u_i is assumed iid Normal and independent of the overall error e_{it}. The model specifies an intercept for each office property around an overall mean c_0. The latent effect is measured by u_i which is the difference between the firm intercept b_{0i} and the overall mean c_0. The test of whether the latent effect is significant is a test on the significance of u_i.

Random Effect Across Office Properties (i) and Across Office Area (j)

This model can be represented by

(7)

$$\ln(y_{it}) = b_{0i} + b_{1j} \ln(x_{1it}) + \cdots + b_M \ln(x_{Mit}) + \ln(a_1 z_{1it} + \cdots + a_A z_{Ait}) + e_{it}$$

$$b_{0i} = c_0 + u_i$$
$$b_{1j} = c_1 + v_j$$

where the errors u_i and v_j are assumed iid Normal and independent of the overall error e_{it}. The model specifies an intercept for each office property around an overall mean c_0 and a slope effect across office areas around the mean c_1. The subscript j in y, x, z, and e, denoting office area j, is suppressed for brevity of notation. The latent effects are the deviations from the overall means and are measured by the firm effect u_i and the office area effect v_j. The tests for the presence of latent effects are therefore tests on the significance of the errors u_i and v_j.

Random Effect Across Office Properties (i) and Across Office Class (k)

This model can be presented as follows:

(8)

$$\ln(y_{it}) = b_{0i} + b_1 \ln(x_{1it}) + b_{2k} \ln(x_{2it}) + \cdots + b_M \ln(x_{Mit}) + \ln(a_1 z_{1it} + \cdots + a_A z_{Ait}) + e_{it}$$

$$b_{0i} = c_0 + u_i$$
$$b_{2k} = c_2 + w_k$$

where the errors u_i and w_k are assumed iid Normal and independent of the overall error e_{it}. The model specifies an intercept for each office property around an overall mean c_0 and a slope effect across office class around the mean c_2. The subscript k in y, x, z, and e, denoting office class k, is suppressed for brevity of notation. The latent effects are the deviations from the overall means and are measured by the firm effect u_i and the office class effect w_k. The tests for the presence of latent effects are therefore tests on the significance of the errors u_i and w_k.

Random Effect Across Office Properties (i) in the Additive Specification (Mean Group Effect on y_{it})

This model is specified as

(9)

$$\ln(y_{it}) = b_0 + b_1 \ln(x_{1it}) + b_2 \ln(x_{2it}) + \cdots + b_M \ln(x_{Mit}) + \ln(a_{0i} + a_1 z_{1it} + \cdots + a_A z_{Ait}) + e_{it}$$

$$a_{0i} = d_0 + r_i$$

where the error r_i is assumed iid Normal and independent of the overall error e_{it}. The model specifies a change in the mean of y_{it} for each office property around an

overall mean d_0. The latent effect is the deviation from the overall mean and is measured by the firm effect r_i. The test of whether the latent effect is significant is a test on the significance of r_i.

The difference between equation 6 and equation 9 is that equation 6 captures the intercept effect in the log regression. This translates to a group effect on the multiplier on y_{it}. It is a *scale* effect, in contrast to equation 9 which is the *level* effect on y_{it}.

Modeling Income

We illustrate our methodology by modeling income and expense for office buildings in Manhattan. Each year the Department of Finance receives filings from these properties and develops models that estimate income and expense for the whole population of filers and nonfilers. The filers are used to develop equations (in-sample); then these equations are used to predict estimates for nonfilers (out-of-sample). This section discusses the formulation of the income equation. The discussion of out-of-sample estimation is presented later in the article.

Data: Manhattan Office Filers

There are 2,294 office buildings in Manhattan of which 1,552 were valid filers in 2013. To create a panel dataset, we use three years of historical data. Of the 1,552 filers in 2013, there are 1,480 that also filed in years 2012, 2011, and 2010. Therefore, the sample size we start with is 1,480 filers over a four-year period.

Outliers, which we define as extreme changes in year-over-year income beyond the 1% tails on each end and expense ratios outside of 10% to 90%, were removed. Thus of the 1,480 filers each period, the estimating sample is 1,266 after outlier removal. Figure 1 maps the location of the 1,266 data points. The majority of office properties are concentrated in the Midtown and Downtown sections of Manhattan with other areas showing sparse data.

Office buildings in New York City are grouped into classes that reflect their quality and the amenities offered. Office classes in Manhattan are identified as trophy, class A, class B, and nontrophy/A/B. Table 1 shows the breakdown of income, expenses, net operating income (NOI), and expense ratios by office class for 2013 filings. Of note is the wide range in income levels—for example, the minimum class A rent is $28.57 psf with a maximum of $143.06 psf. Even when office class is broken down by office area (table 2), the spread in income and expense levels is wide. The Plaza, where the majority of trophy buildings are located, has a minimum rent of $54.49 psf and a maximum rent of $153.30 psf in its premier trophy properties.

The model specification must take into account the variation in income and expense across office class and office area, plus additional variation not explained by these groupings. To test our models, we randomly sample the 1,266 points for a hold-out and estimating sample. The

Figure 1. Map of sample data for Manhattan office filers, 2010–2013

Table 1. Data summary of income, expense, NOI, and expense ratios for Manhattan office filers 2013—Breakdown by Office Class

Office Class	Number of observations	Variable	Minimum	50th percentile	Maximum	Standard deviation
Trophy	38	income	$44.07	$78.27	$153.30	$24.86
		expense	$12.83	$21.10	$37.61	$4.57
		NOI	$28.34	$54.67	$125.88	$21.76
		expense ratio	18%	28%	39%	5%
Class A	186	income	$28.57	$61.10	$143.06	$20.84
		expense	$5.55	$18.69	$36.04	$4.54
		NOI	$17.63	$42.70	$113.49	$17.21
		expense ratio	15%	30%	40%	5%
Class B	345	income	$15.62	$45.06	$118.67	$13.76
		expense	$5.44	$14.50	$37.69	$3.66
		NOI	$10.18	$30.62	$92.88	$11.35
		expense ratio	14%	33%	43%	5%
Nontrophy/A/B	697	income	$9.93	$41.29	$269.84	$22.81
		expense	$2.82	$11.18	$57.78	$5.43
		NOI	$5.83	$29.87	$212.06	$18.83
		expense ratio	10%	28%	48%	7%

Table 2. Data summary of income, expense, NOI, and expense ratios for Manhattan office filers 2013—Breakdown by prime office areas and office class

Office Area	Office Class	Number of observations	Variable	Minimum	50th percentile	Maximum	Standard deviation
Financial/ World Trade	Trophy	4	income	$44.07	$54.18	$61.60	$7.22
			expense	$15.73	$18.12	$20.62	$2.36
			NOI	$28.34	$36.07	$40.97	$5.55
			expense ratio	30%	35%	37%	3%
	Class A	12	income	$35.22	$42.41	$54.64	$6.69
			expense	$9.33	$15.60	$19.58	$2.72
			NOI	$21.90	$28.16	$36.99	$4.69
			expense ratio	25%	37%	38%	4%
	Class B	26	income	$29.50	$36.30	$56.46	$7.73
			expense	$9.74	$14.45	$20.41	$2.54
			NOI	$17.58	$22.01	$38.60	$5.41
			expense ratio	32%	39%	42%	3%
	Nontrophy/A/B	14	income	$26.33	$39.67	$64.00	$11.81
			expense	$6.71	$12.53	$16.17	$2.98
			NOI	$15.45	$26.29	$49.28	$10.06
			expense ratio	20%	32%	42%	6%
Grand Central	Trophy	6	income	$63.49	$84.52	$115.79	$17.79
			expense	$17.78	$19.37	$23.22	$2.21
			NOI	$45.65	$64.01	$93.94	$16.87
			expense ratio	19%	24%	28%	4%

(continued on page 44)

Table 2. Data summary of income, expense, NOI, and expense ratios for Manhattan office filers 2013—Breakdown by prime office areas and office class (*continued*)

Office Area	Office Class	Number of observations	Variable	Minimum	50th percentile	Maximum	Standard deviation
	Class A	43	income	$41.79	$60.65	$98.66	$10.26
			expense	$11.58	$18.70	$23.12	$2.88
			NOI	$29.04	$40.90	$76.15	$9.21
			expense ratio	19%	32%	40%	5%
	Class B	27	income	$15.62	$50.29	$70.53	$12.64
			expense	$5.44	$17.03	$22.68	$3.96
			NOI	$10.18	$34.25	$48.35	$8.99
			expense ratio	25%	34%	35%	3%
	Nontrophy/A/B	28	income	$23.23	$51.67	$95.68	$18.79
			expense	$9.36	$16.46	$39.53	$6.05
			NOI	$13.63	$33.83	$70.23	$15.29
			expense ratio	13%	35%	46%	8%
Insurance/ Civic Center	Trophy	1	income	$52.66	$52.66	$52.66	-
			expense	$12.83	$12.83	$12.83	-
			NOI	$39.83	$39.83	$39.83	-
			expense ratio	24%	24%	24%	-
	Class A	10	income	$28.57	$38.07	$48.81	$5.90
			expense	$5.55	$13.34	$18.35	$3.26
			NOI	$17.63	$24.75	$33.36	$5.09
			expense ratio	15%	38%	38%	8%
	Class B	27	income	$23.34	$37.91	$100.42	$14.72
			expense	$9.59	$14.97	$37.69	$5.30
			NOI	$13.59	$22.72	$62.73	$9.55
			expense ratio	32%	39%	42%	3%
	Nontrophy/A/B	59	income	$16.77	$40.36	$104.23	$15.80
			expense	$4.58	$10.91	$26.06	$4.65
			NOI	$11.85	$28.10	$78.17	$12.52
			expense ratio	14%	28%	43%	8%
Plaza	Trophy	24	income	$54.49	$83.82	$153.30	$26.12
			expense	$18.03	$21.48	$37.61	$4.58
			NOI	$33.24	$62.14	$125.88	$22.81
			expense ratio	18%	28%	39%	5%
	Class A	82	income	$35.99	$70.21	$143.06	$23.29
			expense	$10.94	$20.56	$36.04	$4.84
			NOI	$25.05	$48.60	$113.49	$19.20
			expense ratio	19%	30%	37%	4%
	Class B	15	income	$36.52	$51.44	$77.86	$10.95
			expense	$8.82	$16.75	$20.85	$3.00
			NOI	$23.80	$34.18	$58.31	$8.78
			expense ratio	21%	32%	35%	4%
	Nontrophy/A/B	52	income	$29.85	$60.40	$269.84	$36.35
			expense	$8.28	$18.01	$57.78	$7.63
			NOI	$19.60	$42.52	$212.06	$29.93
			expense ratio	14%	27%	45%	7%

hold-out sample has 254 data points, or 20% of 1,266, and our estimating sample has the remaining 1,012 data points. Table 3 shows the characteristics of the full, estimating, and hold-out samples. The means, medians, and standard deviation of income, expense, NOI, and expense ratios are similar. The estimating sample with 1,012 filers is used throughout this section.

Model Specification Using Single-year Data

The selection of variables entered into the model is accomplished using 2013 data, the most recent year in our dataset. We work with equation 4 for this purpose. The dependent variable is the natural log of income per square foot, denoted as $\ln(incpsf)$. The independent variables are selected using stepwise regression, retaining the group effects office area and office class. The office area and office class variables are retained through the elimination process of the stepwise procedure because, as indicated in the previous section, there is heterogeneity within these groups that needs to be modeled. If heterogeneity within these groups is not taken into account, certain variables can turn insignificant in a stepwise procedure due to unexplained variation. The model estimated is

$$\ln(incpsf) = \beta_0 + \beta_1 \ln(Office\ Area) + \beta_2 \ln(Office\ Class) + \ln(\alpha_1 altage + \alpha_2 offsqft + \alpha_3 offratio + \alpha_4 retratio + \alpha_5 resratio + \alpha_6 landsize + \alpha_7 story) + e$$

where

Office Area is the office area adjustments,

Office Class is the office class adjustments,

altage is the age since altered,

offsqft is the office area square footage,

offratio is the percentage of the total gross square footage that is office area,

Table 3. Income and expense distribution for full sample, estimating sample, and hold-out sample for Manhattan office filers 2013

Full Sample				
Variable	**Number**	**Median**	**Mean**	**Standard deviation**
income	1266	$46.18	$50.89	$22.14
expense	1266	$13.54	$14.20	$5.52
NOI	1266	$32.25	$36.69	$18.01
expense ratio	1266	29%	29%	7%

Estimating Sample				
Variable	**Number**	**Median**	**Mean**	**Standard deviation**
income	1012	$45.92	$50.85	$22.24
expense	1012	$13.54	$14.19	$5.50
NOI	1012	$31.94	$36.66	$18.09
expense ratio	1012	30%	29%	7%

Hold-out Sample				
Variable	**Number**	**Median**	**Mean**	**Standard deviation**
income	254	$47.20	$51.05	$21.78
expense	254	$13.54	$14.25	$5.62
NOI	254	$32.65	$36.79	$17.71
expense ratio	254	29%	29%	6%

retratio is the percentage of the total gross square footage that is retail area,

resratio is the percentage of the total gross square footage that is residential area,

landsize is the land size in square feet,

story is the number of stories for the entire building.

The results of the nonlinear regression are given in figure 2. The summary statistics for model performance are as follows:

- *Error ratio.* This statistic indicates the predicted income to the actual income; therefore a number closer to 1 is indicative of a better model. In this model, the median error ratio is 1.024.
- *Pseudo* R^2. This measure is calculated by taking the square of the correlation of the predicted and actual income. It has the same interpretation as the R^2 in linear regressions. The pseudo R^2 for this model is not strong at 52.07%.
- *Coefficient of dispersion* (COD). This statistic measures the spread around the median error ratio. It is calculated as the weighted dispersion of the error ratio of each parcel from the median error ratio. A COD between 5 and 15 is considered indicative of a good model. The COD of this model is 27.39 indicating an error spread that is wider than is normally acceptable.
- *Regressivity/progressivity.* This measure serves a similar function to the price-related differential (PRD). The PRD typically is applied to sale prices; however, the concept and calculation can be extended to

Figure 2. Income regression model for Manhattan office filers 2013

Summary Statistics

(Sample size = 1012)

Median error ratio	1.024
Pseudo R^2	52.07%
COD	27.39
Regressivity	1.106

Nonlinear Regression

$$\ln(incpsf) = \beta_0 + \beta_1 \ln(Office\ Area) + \beta_2 \ln(Office\ Class) + \ln(\alpha_1 altage + \alpha_2 offsqft + \alpha_3 offratio + \alpha_4 retratio + \alpha_5 resratio + \alpha_6 landsize + \alpha_7 story) + e$$

Dependent variable: ln(Income Per Square Foot)

Variable	Parameter estimate	Standard error	*t*-value	Pr > |*t*|
Intercept	−0.19060	0.01662	−11.47	<.0001
Office Area	0.86230	0.07097	12.15	<.0001
Office Class	−0.02868	0.08845	−0.32	0.7458
altage	−0.06126	0.02182	−2.81	0.0051
offsqft	−0.00002	0.00001	−3.44	0.0006
offratio	36.65070	2.67140	13.72	<.0001
retratio	110.67000	6.79710	16.28	<.0001
resratio	30.35600	12.66850	2.40	0.0167
landsize	0.00038	0.00008	4.76	<.0001
story	0.72570	0.11090	6.55	<.0001

Office Class Adjustments	
Trophy	1.724
Class A	1.399
Class B	**1.000**
Nontrophy/A/B	0.952

Office Area Adjustments	
Financial/World Trade	**1.000**
Grand Central	1.523
Grand Central South	1.210
Harlem East	0.760
Insurance/Civic Center	0.997
Manhattan–North 110th	0.858
Midtown South	1.097
Midtown West	1.041
Plaza	1.696
Upper East Side (59th–96th)	2.056
Upper West Side (59th–96th)	1.564

any variable. The regressivity/progressivity statistic measures how the error ratio moves in relation to the *level* of the variable being estimated. For our income variable, if the error ratio decreases for higher levels of income, then we call this regressive (we are underestimating income for higher-income properties). Similarly, if the error ratio increases as the income level increases, then the model is progressive (we are overestimating income for higher-income properties). The desirable range is between 0.98 and 1.03, with values higher than 1.03 indicating significant regressivity. Our model is regressive with a measure of 1.106.

The group effects specified in the model—the *Office Area* and *Office Class* variables—are multiplicative adjustments to the physical characteristics. Office class adjustments are calculated as the ratio of the median income of each class to the class B median. An office class with a ratio greater than 1 means an upward adjustment is warranted. For example, as shown in figure 2, the trophy class has a ratio of 1.724; this means trophy properties get a 72.4% upward adjustment in rents, while properties in class B get no adjustment (the ratio is 1). Similarly, office area adjustments are constructed as ratios of median rents by area to a base median. In this case, the base is the median rent in the Financial/World Trade area. Thus, in the Plaza area, with an adjustment factor of 1.696, rents are adjusted upward by 69.6%.

These results indicate that the two group effects, meant to address heterogeneity across office properties, are inadequate since the model does not significantly explain the variation in income. Thus, unmeasured effects that we know affect income must be included in the error. For example, we know that income is strongly related to historical income because leases often contain provisions for rent escalation. Table 4 shows the correlation of income filed in 2013 and historical filings. The correlation of income filed in 2013 with 2012 is 95.73%, the correlation with 2011 filings is 92.70%, and with 2010 filings, the correlation is 90.32%. The use of historical income can therefore improve prediction by modeling the unmeasured effects in the residuals.

Table 4. Correlation of income per square foot between years 2013, 2012, 2011, and 2010

Pearson Correlation Coefficients, N = 1012 Prob > |r| under H0: Rho=0				
	Income 2013	Income 2012	Income 2011	Income 2010
Income 2013	1	0.95730	0.92703	0.9032
		<.0001	<.0001	<.0001
Income 2012	0.95730	1	0.97148	0.94327
	<.0001		<.0001	<.0001
Income 2011	0.92703	0.97148	1	0.96271
	<.0001	<.0001		<.0001
Income 2010	0.90320	0.94327	0.96271	1
	<.0001	<.0001	<.0001	

Modeling Latent Effects: Panel Data Models

We include historical information in the prediction of income in a panel data structure. The sample size for this section is 4,048 which consist of the 1,012 filers in each of the years 2013, 2012, 2011, and 2010. We ran five models—one pooled regression and four random effects regressions—based on the framework introduced previously.

Pooled Regression

A pooled regression based on equation 5 was run with the variables selected for the single-year data model plus an additional variable, the inflation rate per period. This variable was added since rent escalations are usually indexed to the Consumer Price Index (CPI). The

inflation rate is calculated as the year-over-year change in the CPI. The model estimated is

$$\ln(incpsf_{it}) = b_0 + b_1\ln(Office\ Area_{it}) + b_2\ln(Office\ Class_{it}) + b_3\ln(CPI\ Rate_{it}) + \ln(a_1 altage_{it} + a_2 offsqft_{it} + a_3 offratio_{it} + a_4 retratio_{it} + a_5 resratio_{it} + a_6 landsize_{it} + a_7 story_{it}) + e_{it}.$$

The results are presented in appendix A, figure A1. As expected, the pooled regression performs like the single-year data model because there is no attempt to explain group effects other than by office area and office class. The median error ratio is 1.026, the pseudo R^2 is 51.86%, the COD is 28.71, and the regressivity measure is 1.104. These summary model performance measures are not significantly different than those based on single-year data. Thus, the pooled regression specification does not capture the additional information that historical information can provide.

Random Effects Regression

Random Effect Across Office Properties (i)
This model, which is based on equation 6, is written as

$$\ln(incpsf_{it}) = b_{0i} + b_1\ln(Office\ Area_{it}) + b_2\ln(Office\ Class_{it}) + b_3\ln(CPI\ Rate_{it}) + \ln(a_1 altage_{it} + a_2 offsqft_{it} + a_3 offratio_{it} + a_4 retratio_{it} + a_5 resratio_{it} + a_6 landsize_{it} + a_7 story_{it}) + e_{it}.$$

$$b_{0i} = c_0 + u_i.$$

Heterogeneity is now modeled by the intercepts varying between office properties. The effect differentiates the scale adjustments across offices. As shown in the summary statistics table in appendix A, figure A2, the improvement in the model is significant with a median error ratio of 0.997, a pseudo R^2 of 89.77% (up from 51.86%), a COD of 9.53 (down from 28.71), and a regressivity measure of 1.014, which is below 1.03 indicating there is no regressivity in the estimates.

The improvement in the model fit results from the specification of the second equation, $b_{0i} = c_0 + u_i$, which is statistically significant as shown in the last table of figure A2 in appendix A. This table presents the *t*-values that test the significance of the change in intercept by firm. An absolute value of the *t*-ratios greater than 1.96 indicates a significant deviation of the firm intercept from the overall mean c_0 at the 95% confidence interval. More than 50% of the office properties have significantly different intercepts from c_0.

Random Effect Across Office Properties (i) and Across Office Area (j)
The model, which is based on equation 7, is written as

$$\ln(incpsf_{it}) = b_{0i} + b_{1j}\ln(Office\ Area_{it}) + b_2\ln(Office\ Class_{it}) + b_3\ln(CPI\ Rate_{it}) + \ln(a_1 altage_{it} + a_2 offsqft_{it} + a_3 offratio_{it} + a_4 retratio_{it} + a_5 resratio_{it} + a_6 landsize_{it} + a_7 story_{it}) + e_{it}.$$

$$b_{0i} = c_0 + u_i$$
$$b_{1j} = c_1 + v_j.$$

Heterogeneity is now modeled by the intercepts varying between office properties and slopes varying across office areas. The addition of the varying slope by office area is not significant. The summary performance measures (appendix A, figure A3) do not improve on the firm-varying intercept in the previous model (appendix A, figure A2). The median error ratio is 0.996, the pseudo R^2 is 89.85% (up from 89.77%), the COD is 9.50 (down from 9.53), and the regressivity measure is the same at 1.014 which indicates no significant regressivity in estimates.

The last table in figure A3 in appendix A presents the tests of significance of the second equation, $b_{1j} = c_1 + v_j$. Except for Midtown West, the *t*-values indicate that office area effects as modeled by different slopes are not significant.

Random Effect Across Office Properties (i) and Across Office Class (k)

This model, which is based on equation 8, is written as

$$\ln(incpsf_{it}) = b_{0i} + b_1\ln(Office\ Area_{it}) + b_{2k}\ln(Office\ Class_{it}) + b_3\ln(CPI\ Rate_{it}) + \ln(a_1 altage_{it} + a_2 offsqft_{it} + a_3 offratio_{it} + a_4 retratio_{it} + a_5 resratio_{it} + a_6 landsize_{it} + a_7 story_{it}) + e_{it}$$

$$b_{0i} = c_0 + u_i$$
$$b_{2k} = c_2 + w_k.$$

Heterogeneity is now modeled by the intercepts varying between office properties and slopes varying across office class. The addition of the varying slope by office class is not significant. The summary performance measures (appendix A, figure A4) do not improve on the firm-varying intercept in the first random effects model (appendix A, figure A2). The median error ratio is 0.998, the pseudo R^2 is 90.02% (up from 89.77%), the COD is 9.46 (down from 9.53), and the regressivity measure is the same at 1.014 which indicates no significant regressivity in estimates.

The last table in figure A4 in appendix A shows the results of significance tests of the equation, $b_{2k} = c_2 + w_k$. None of the *t*-values show that office area effects as modeled by different slopes are significant.

Random Effect Across Office Properties (i)—Level Effects on Income

This model, which is based on equation 9, is written as

$$\ln(incpsf_{it}) = b_0 + b_1\ln(Office\ Area_{it}) + b_2\ln(Office\ Class_{it}) + b_3\ln(CPI\ Rate_{it}) + \ln(a_{0i} + a_1 altage_{it} + a_2 offratio_{it} + a_3 retratio_{it} + a_4 resratio_{it} + a_5 landsize_{it} + a_6 story_{it}) + e_{it}$$

$$a_{0i} = d_0 + r_i.$$

The results are reported in appendix A, figure A5. Heterogeneity is now modeled by the intercepts that affect the income level varying between office properties. The median error ratio is 0.997, the pseudo R^2 is 95.74% (up from 89.77%), the COD is 6.13 (down from 9.53), and the regressivity measure is lower at 1.009 which indicates no significant regressivity in estimates.

The improvement in the model fit can be attributed to the specification of the equation, $a_{0i} = d_0 + r_i$, which is statistically significant as shown in the last table of figure A5 in appendix A. The median of the *t*-value ratios is −1.4080, which, compared to −0.2186 in appendix A, figure A2, represents more intercepts that significantly vary from the overall mean. Heterogeneity is better captured by varying the mean level rather than the scale of income across office properties.

Final Model Selection

Based on the previous results, the final model for income is the level effects model. Thus, the predicted income for office property *i* in filing period *t* is calculated as

(10)

$$\widehat{incpsf}_{it} = e^{-0.074}(Office\ Area_{it})^{0.022} (Office\ Class_{it})^{0.009}(CPI\ Rate_{it})^{2.451} (31.52 + \hat{r}_i + 0.041\ altage_{it} + 4.584\ offratio_{it} + 27.792\ retratio_{it} + 0.0001\ landsize_{it} + 0.328\ story_{it})$$

where $\hat{r}_i$ is the estimated latent effect. The estimated latent effect is now part of the measurement equation which includes the observed building and land characteristics (*X* and *Z* in our notation).

Modeling Expenses

Expenses are modeled from expense ratios and income using the identity,

$$Expenses = ExpRatio \times Income.$$

Given estimated income using equation 10 and estimated expense ratios, the expense model is

(11)

$$\ln(exppsf_{it}) = g_0 + g_1\ln(\widehat{expratio}_{it}) + g_2\ln(\widehat{incpsf}_{it}) + \epsilon_{it}$$

where *exppsf* denotes expense per square foot.

Modeling the Expense Ratio

The steps for modeling expense ratios are the same as for modeling income. A stepwise regression was run on single-year data (2013) to determine significant variables. With the model variables determined, several panel data models were run to determine the best fit. We report only the random effects models as these gave the best results (see appendix B, table B1). The best fit was obtained from the expense-ratio model measuring the level effect varying by office property, with a median error ratio of 0.995, a pseudo R^2 of 80.75%, a COD of 12.71, and a regressivity measure of 1.039.

The estimated model is given in appendix B, figure B1 together with the *t*-value tests of whether the changing intercepts are significant. The median error ratio is 0.9947, the pseudo R^2 is 80.75%, the COD is 12.71, and the regressivity measure is 1.039. The *t*-statistics for testing the significance of the different intercepts are significant for over 50% of the properties. From the figure B1 model in appendix B, the predicted expense ratio for office property in filing period *t* is calculated as

(12)

$$\widehat{expratio}_{it} = e^{0.0184}(Office\ Area_{it})^{-0.015}(Office\ Class_{it})^{0.022}(CPI\ Rate_{it})^{-1.5596}(0.32 + \tilde{r}_i + 0.00009\ altage_{it} + 0.02898\ offratio_{it} - 0.125\ retratio_{it} - 0.05\ resratio_{it} + (9.36E-06)\ bldfrt_{it} + 0.0012\ story_{it})$$

where *bldfrt* is building frontage in feet and $\tilde{r}_i$ is the estimated latent effect.

Modeling Expense

Expense is then estimated using estimated income (from equation 10) and the estimated expense ratio (from equation 12). This section compares two expense model specifications—a pooled regression model where parameters are constant and a random coefficients model that varies by time period.

Pooled Regression

A linear regression on equation 11 is a pooled regression with the parameters (g_0, g_1, and g_2) constant across groups. It should be noted, however, that in this equation, firm (or office property) heterogeneity is captured by $\widehat{incpsf}_{it}$ and $\widehat{expratio}_{it}$. Figure B2 in appendix B contains the results of the pooled regression. The median error ratio is 0.998, the pseudo R^2 is 88.84%, the COD is 13.89, and the regressivity measure is 1.035.

Random Effects Regression—Time-varying Coefficients

Heterogeneity across firms is captured through our models of income and expense ratios. It could be, however, that the relationship of expense and income is different across filing periods. This means that the regression (equation 11) should be run for different filing periods instead of as a panel. Time-varying coefficients can be tested by estimating the model,

(13)

$$\ln(exppsf_{it}) = g_{0t} + g_{1t}\ln(\widehat{expratio}_{it}) + g_{2t}\ln(\widehat{incpsf}_{it}) + \epsilon_{it}$$

$$g_{0t} = f_0 + \gamma_{0t}$$
$$g_{1t} = f_1 + \gamma_{1t}$$
$$g_{2t} = f_2 + \gamma_{2t}$$

where γ_{0t}, γ_{1t}, and γ_{2t} are assumed iid Normal and independent of the overall error ϵ_{it}. The model specifies different intercepts and slopes for each time period implying a time-varying relationship between expense and income.

Figure B3 in appendix B contains the results of the time-varying coefficients regression. The performance improves slightly with the median error ratio at 0.9926, the pseudo R^2 at 90.80% (up from 88.84%), the COD at 13.01 (down from 13.89), and the regressivity measure at 1.0294 (down from 1.035).

The last table in figure B3, appendix B shows that filing periods 2011 and

2013 have significantly different intercepts and slopes. Therefore, the model for expense should be the time-varying coefficients model. The 2013 regression coefficients are derived as

$$g_0 = 0.2529 - 0.1178 = 0.1351$$
$$g_1 = 1.2341 - 0.2348 = 0.9993$$
$$g_2 = 1.0075 - 0.0485 = 0.9590.$$

Income and Expense Estimates

Finally, income and expense estimates for the period of interest, in this case 2013, are obtained by the following equations (using $t = 2013$ data):

Income Model Specification:

$$\widehat{incpsf}_i = e^{-0.074}\ (Office\ Area_i)^{0.022}\ (Office\ Class_i)^{0.009}(CPI\ Rate_i)^{2.451}\ (31.52 + \hat{r}_i + 0.041\ altage_i + 4.584\ offratio_i + 27.792\ retratio_i + 0.0001\ landsize_i + 0.328\ story_i)$$

Expense Ratio Model Specification:

$$\widehat{exprati}o_i = e^{0.0184}\ (Office\ Area_i)^{-0.015}\ (Office\ Class_i)^{0.022}(CPI\ Rate_i)^{-1.5596}\ (0.32 + \tilde{r}_i + 0.00009\ altage_i - 0.02898\ offratio_i - 0.125\ retratio_i - 0.05\ resratio_i + (9.36E-06)\ bldfrt_i + 0.0012 story_i)$$

Expense Model Specification:

$$\widehat{exppsf}_i = e^{0.1351}(\widehat{exprati}o_i)^{0.9993}(\widehat{incpsf}_i)^{0.9590}$$

Figure B4 in appendix B compares the actual and predicted income, expense, and NOI for 2013. The first table shows the distribution of the actual and predicted values. The model predicts the actual values well even in the 1% and 99% tails. The fit at the median and mean are tight. The graphs show the predicted estimates on the *y*-axis and actual values on the *x*-axis. Comparing the results here to the 2013 income estimates without the panel data structure (figure 2), the median income error ratio is 1.002 (down from 1.024), the pseudo R^2 is 96.90% (up from 52.07%), the COD is 6.58 (down from 27.39), and the regressivity measure indicates no regressive estimates at 1.011 (down from 1.106). These improvements highlight the importance of modeling unobserved heterogeneity which otherwise would remain unmeasured in the residuals.

Extension to Nonfilers

The methods used to calculate income, expense ratios, and expenses for filers can be applied to nonfilers as well. Each year estimates of income and expense must be determined for all offices, whether they are filers or nonfilers. Nonfilers are considered an out-of-sample population since they have no income or expense to include as a dependent variable in the regressions.

The income and expense ratio equations previously described enable us to enter in the physical attributes of nonfilers and obtain estimates of income, expense ratios, expenses, and hence NOI which when capitalized equals value. The only attribute that nonfilers do not have are the estimated latent effects $\hat{r}_i$ in the income equation and $\tilde{r}_i$ in the expense ratio equation. Each firm in our filer dataset has an estimated latent effect and *x*-*y* coordinates that denote its location. From this information, a three-dimensional graph of the latent effect (*z*-axis) on the *x*-*y* coordinates can be rendered that shows the geographical (or spatial) distribution. This relationship between the latent effects and the *x*-*y* coordinates, which we refer to as a surface, can be estimated by several known methods such as spline, kriging, k-nearest neighbor, and inverse weighted distance. In this way, any parcel that has an *x*-*y* coordinate can be mapped on this estimated surface to obtain the interpolated latent effect.

In summary, the steps to obtain estimates for the nonfiler population are: first, develop estimating equations for income, expense ratios, and expense from filings as outlined in the Estimating Income and Modeling Expenses sections; second, interpolate latent effect estimates for nonfilers using the surface generated by filers; and third, predict out-of-sample

estimates for nonfilers by entering the interpolated latent effect and physical characteristics into the three equations.

Verification of the performance of the out-of-sample predications is done by applying the estimating equations used for nonfilers to a hold-out sample of filers. In this case, the hold-out sample is the 254 filers that were not included in the derivation of the equations. By holding out a sample of filers and treating them as we would nonfilers, we can compare the estimates from the holdout sample to their actual filings to gauge the out-of-sample model performance.

Figure C1 in appendix C summarizes these steps for our hold-out sample. The estimated latent effects for income, $\hat{r}_i$, and for expense ratios, $\tilde{r}_i$, are graphed as surfaces on the *x-y* coordinates. Both the income and expense-ratio latent effect surfaces are estimated using spline. Each surface shows the geographical distribution of the latent effects. It should be noted that the surfaces are continuous, i.e., the latent effects are interpolated for all *x-y* coordinates. Thus, for any office property in Manhattan, we can estimate its income and expense-ratio latent effect based on its location, i.e., its *x-y* coordinate. Therefore, given the location (*x-y* coordinate) of any property in the hold-out sample, the surface will tell us its estimated latent effect.

By entering the interpolated latent effect and physical characteristics of hold-out sample properties into the estimating equations, we can obtain the income, expense ratio, and expense estimates. These estimates are compared to the actual filings in the hold-out sample to get the summary measures of model performance shown in appendix C, figure C1. For income, the median error ratio is 1.0117, the pseudo R^2 is 66%, the COD is 26.23, and the regressivity is 1.1033.

Table C1 in appendix C compares the summary measures shown in figure C1 for hold-out sample models with latent effects to models without latent effects. The model improvement with latent effects is evident in the expense ratio and expense models. For expense ratios, the pseudo R^2 increases to 52.02% (from 26.53%) and the COD decreases to 17.67 (from 20.08). For expenses, the pseudo R^2 increases to 67.09% (from 55.21%), the COD decreases to 25.36 (from 31.91), and the regressivity measure decreases to 1.0979 (from 1.139). Thus, based on this in-sample performance, it can be demonstrated that when applied to the out-of-sample estimates of nonfilers, the latent variable models will perform better than models that do not estimate latent effects.

Summary

In summary, this article introduced a method of improving annual regression estimates of income and expense by using historical information to extract latent effects. Historical information is incorporated in a panel data structure which has time and cross-section dimensions. The panel structure gives access to the class of latent variable models that estimate and test the presence of group effects (heterogeneity), which are unobserved but known to affect income or expense. Heterogeneity in the panel data structure manifests itself in different intercepts and/or slopes across groups. In our case, intercepts were allowed to vary for each office property giving us a distribution of mean income and mean expense ratios. Using this distribution, we estimated a surface across *x-y* coordinates using spline. By estimating this surface, we are able to interpolate latent effects for any property—even those without income and expense filings—based on location. We showed that incorporating latent effects increases performance both in-sample and out-of-sample. Thus this article not only contributes to the literature in the use of latent variable (panel data) models to estimate income and expense, but also extends it by illustrating an out-of-sample procedure, based on interpolation

methods, to predict latent effects. We showed that inclusion of latent effect estimates substantially improves income and expense predictions for both filers and nonfilers.

References

Baltagi, B. 2008. *Econometric analysis of panel data*, 4th ed. Hoboken, NJ: John Wiley & Sons.

Francke, M.K. 2008. The hierarchical trend model. In *Mass appraisal methods: An international perspective for property valuers*, ed. T. Kauko and M. d'Amato, 164–180. Oxford: Wiley-Blackwell.

Francke, M.K., and G.A. Vos. 2004. The hierarchical trend model for property valuation and local price indices. *Journal of Real Estate Finance and Economics* 28 (2/3): 179–208.

Gloudemans, R.J. 1999. *Mass appraisal of real property*. Chicago: International Association of Assessing Officers.

Hsiao, C. 2003. *Analysis of panel data*, 2nd ed. New York: Cambridge University Press.

IAAO. 1990. *Property appraisal and assessment administration*. Chicago: International Association of Assessing Officers.

IAAO. 2003. *Standard on automated valuation models*. Chicago: International Association of Assessing Officers.

O'Connor, P.M., J.G. Panos, and S. Som. 1991. Manhattan commercial property: Development and comparison of valuation models. *Property Tax Journal* 10 (1): 3–23.

Appendix A. Results from Tests of Five Different Specifications for a Panel Data Income Regression Model

Figure A1. Panel income regression model—Pooled

Summary Statistics
(Sample size = 4048)

Median error ratio	1.026
Pseudo R^2	51.86%
COD	28.71
Regressivity	1.104

Model

$$\ln(incpsf_{it}) = b_0 + b_1\ln(Office\ Area_{it}) + b_2\ln(Office\ Class_{it}) + b_3\ln(CPI\ Rate_{it}) + \ln(a_1 altage_{it} + a_2 offsqft_{it} + a_3 offratio_{it} + a_4 retratio_{it} + a_5 resratio_{it} + a_6 landsize_{it} + a_7 story_{it}) + e_{it}$$

Dependent variable: ln(Income Per Square Foot)

Variable	Parameter estimate	Standard error	*t*-value	Pr > |*t*|
Intercept	−0.14	0.01	−12.66	<.0001
Office Area	0.70	0.03	21.83	<.0001
Office Class	−0.03	0.04	−0.90	0.3660
CPI_Rate	0.27	0.32	0.83	0.4056
altage	0.01	0.01	0.80	0.4214
offsqft	−3.00E-05	2.84E-06	−10.10	<.0001
offratio	26.00	0.97	26.89	<.0001
retratio	105.01	3.18	33.05	<.0001
resratio	21.65	6.03	3.59	0.0003
landsize	0.00	0.00	11.90	<.0001
story	0.93	0.05	19.78	<.0001

Figure A2. Panel income regression model—Office scale effects

Summary Statistics
(Sample size = 4048)

Median error ratio	0.997
Pseudo R^2	89.77%
COD	9.53
Regressivity	1.014

Model

$$\ln(incpsf_{it}) = b_{0i} + b_1 \ln(Office\ Area_{it}) + b_2 \ln(Office\ Class_{it}) + b_3 \ln(CPI\ Rate_{it}) + \ln(a_1 altage_{it} + a_2 offsqft_{it} + a_3 offratio_{it} + a_4 retratio_{it} + a_5 resratio_{it} + a_6 landsize_{it} + a_7 story_{it}) + e_{it}$$

$$b_{0i} = c_0 + u_i$$

Dependent variable: ln(Income Per Square Foot)

Variable	Parameter estimate	Standard error	*t*-value	Pr > |*t*|
Mean Intercept (c_0)	−0.14	0.01	−10.66	<.0001
Office Area	0.69	0.05	13.18	<.0001
Office Class	−0.20	0.05	−4.16	0.3660
CPI_Rate	0.36	0.15	2.36	0.4056
altage	0.01	0.01	0.80	0.4214
offsqft	−3.00E-05	2.84E-06	−10.10	<.0001
offratio	26.00	0.97	26.89	<.0001
retratio	105.01	3.18	33.05	<.0001
resratio	21.65	6.03	3.59	0.0003
landsize	0.00	0.00	11.90	<.0001
story	0.93	0.05	19.78	<.0001

Distribution of Estimates of b_{0i} and *t*-values that Test Whether Difference from Mean c_0 is Significant

Variable	Number	Minimum	25th percentile	50th percentile	75th percentile	Maximum	Mean
b_{0i}	1012	−1.1337	−0.3450	−0.1553	0.0599	1.1843	−0.1370
t-value	1012	−12.1281	−2.4962	−0.2186	2.3806	15.2439	−0.0043

Figure A3. Panel income regression model—Office scale and office area effects

Summary Statistics
(Sample size = 4048)

Median error ratio	0.996
Pseudo R^2	89.85%
COD	9.50
Regressivity	1.014

Model

$$\ln(incpsf_{it}) = b_{0i} + b_{1j}\ln(Office\ Area_{it}) + b_2\ln(Office\ Class_{it}) + b_3\ln(CPI\ Rate_{it}) + \ln(a_1 altage_{it} + a_2 offsqft_{it} + a_3 offratio_{it} + a_4 retratio_{it} + a_5 resratio_{it} + a_6 landsize_{it} + a_7 story_{it}) + e_{it}$$

$$b_{0i} = c_0 + u_i$$
$$b_{1j} = c_1 + v_j$$

Distribution of Estimates of b_{0i} and t-values that Test Whether Difference from Mean c_0 is Significant

Variable	Number	Minimum	25th percentile	50th percentile	75th percentile	Maximum	Mean
b_{0i}	1012	−1.1262	−0.3372	−0.1496	0.0652	1.0977	−0.1329
t-value	1012	−12.0916	−2.4074	−0.1784	2.3054	14.8835	−0.0027

Tests of Whether Difference of b_{1j} from Mean c_1 is Significant (Test for Office Area Effect)

Office Area	Estimate	Std Err Pred	t-value	Pr > |t|
Financial/World Trade	9.17E-11	0.3216	0	1
Grand Central	−0.06748	0.1462	−0.46	0.6445
Grand Central South	−0.09933	0.1966	−0.51	0.6133
Harlem East	0.12280	0.2754	0.45	0.6558
Insurance/Civic Center	0.27770	0.2025	1.37	0.1704
Manhattan—North 110th	−0.05070	0.1852	−0.27	0.7843
Midtown South	−0.13540	0.1712	−0.79	0.4292
Midtown West	−0.58540	0.2131	−2.75	0.0060
Plaza	0.14030	0.1328	1.06	0.2912
Upper East Side (59th–96th)	0.39850	0.1663	2.40	0.0166
Upper West Side (59th–96th)	−0.00097	0.2328	0	0.9967

Figure A4. Panel income regression model—Office scale and office class effects

Summary Statistics
(Sample size = 4048)

Median error ratio	0.998
Pseudo R^2	90.02%
COD	9.46
Regressivity	1.014

Model

$$\ln(incpsf_{it}) = b_{0i} + b_1 \ln(Office\ Area_{it}) + b_{2k} \ln(Office\ Class_{it}) + b_3 \ln(CPI\ Rate_{it}) + \ln(a_1 altage_{it} + a_2 offsqft_{it} + a_3 offratio_{it} + a_4 retratio_{it} + a_5 resratio_{it} + a_6 landsize_{it} + a_7 story_{it}) + e_{it}$$

$$b_{0i} = c_0 + u_i$$
$$b_{2k} = c_2 + w_k$$

Distribution of Estimates of b_{0i} and t-values that Test Whether Difference from Mean c_0 is Significant

Variable	Number	Minimum	25th percentile	50th percentile	75th percentile	Maximum	Mean
b_{0i}	1012	−1.2269	−0.4138	−0.2252	−0.0126	1.1181	−0.2089
t-value	1012	−12.4999	−2.4949	−0.1943	2.3408	15.4490	−0.0041

Tests Whether Difference of b_{2k} from Mean c_2 is Significant (Test for Office Class Effect)

Office Class	Estimate	Std Err Pred	t-value	Pr > |t|
Trophy	0.3425	0.3531	0.97	0.3322
Class A	0.3422	0.3515	0.97	0.3304
Class B	−5.14E-11	0.5994	0	1
Nontrophy/A/B	−0.6846	0.3557	−1.92	0.0544

Figure A5. Panel income regression model—Office level effects

Summary Statistics
(Sample size = 4048)

Median error ratio	0.997
Pseudo R^2	95.74%
COD	6.13
Regressivity	1.009

Model

$$\ln(incpsf_{it}) = b_0 + b_1\ln(Office\ Area_{it}) + b_2\ln(Office\ Class_{it}) + b_3\ln(CPI\ Rate_{it}) + \ln(a_{0i} + a_1 altage_{it} + a_2 offratio_{it} + a_3 retratio_{it} + a_4 resratio_{it} + a_5 landsize_{it} + a_6 story_{it}) + e_{it}$$

$$a_{0i} = d_0 + r_i$$

Dependent variable: ln(Income Per Square Foot)

Variable	Parameter estimate	Standard error	*t*-value	Pr > |*t*|
Intercept	−0.074	0.003	−28.94	<.0001
Office Area	0.022	0.008	2.89	0.0038
Office Class	0.009	0.009	1.11	0.2688
CPI_Rate	2.451	0.076	32.13	<.0001
Mean Base PSF (d_0)	31.522	2.307	13.66	<.0001
altage	0.041	0.021	1.94	0.0521
offratio	4.584	0.644	7.11	<.0001
retratio	27.792	2.920	9.52	<.0001
landsize	0.000	0.000	2.53	0.0113
story	0.328	0.070	4.71	<.0001

Distribution of estimates of a_{0i} and *t*-values that Test Whether Difference from Mean d_0 is Significant

Variable	Number	Minimum	25th percentile	50th percentile	75th percentile	Maximum	Mean
a_{0i}	1012	−0.6926	18.7511	27.1582	39.0586	205.7770	31.5217
t-value	1012	−10.3221	−4.1363	−1.4080	2.4297	56.7492	−0.0279

Appendix B. Results from Tests of Panel Data Expense Regression Model Specifications

Table B1. Expense ratio panel data model comparison

	Office scale effect	Office scale and office area	Office scale and office class	Office level effect
Median error ratio	1.001	1.000	1.003	0.995
Pseudo R^2	74.46%	74.72%	74.75%	80.75%
COD	13.50	13.46	13.41	12.71
Regressivity	1.034	1.033	1.034	1.039

Figure B1. Panel expense ratio regression model—Office level effects

Summary Statistics
(Sample size = 4048)

Median error ratio	0.9947
Pseudo R^2	80.75%
COD	12.71
Regressivity	1.03913

Model

$$\ln(expratio_{it}) = b_0 + b_1\ln(Office\ Area_{it}) + b_2\ln(Office\ Class_{it}) + b_3\ln(CPI\ Rate_{it}) + \ln(a_{0i} + a_1 altage_{it} + a_2 offratio_{it} + a_3 retratio_{it} + a_4 resratio_{it} + a_5 bldfrt_{it} + a_6 story_{it}) + e_{it}$$

$$a_{0i} = d_0 + r_i$$

Dependent variable: ln(Expense Ratio)

Variable	Parameter estimate	Standard error	*t*-value	Pr > |*t*|
Intercept	0.01840	0.004933	3.73	0.0002
Office Area	−0.01471	0.014530	−1.01	0.3113
Office Class	0.02245	0.016460	1.36	0.1726
CPI_Rate	−1.55960	0.146700	−10.63	<.0001
Mean Base (d_0)	0.32200	0.010260	31.39	<.0001
altage	0.00009	0.000078	1.16	0.2470
offratio	−0.02898	0.005976	−4.85	<.0001
retratio	−0.12490	0.021310	−5.86	<.0001
resratio	−0.05446	0.040560	−1.34	0.1794
bldfrt	9.36E-06	0.000011	0.85	0.3945
story	0.001218	0.000235	5.19	<.0001

Distribution of Estimates of a_{0i} and *t*-values that Test Whether Difference from Mean d_0 is Significant

Variable	Number	Minimum	25th percentile	50th percentile	75th percentile	Maximum	Mean
a_{0i}	1012	0.1659	0.2800	0.3207	0.3608	0.5027	0.3220
t-value	1012	−5.3885	−1.4593	−0.0448	1.3545	6.3223	0.0014

Figure B2. Panel expense regression—Pooled

Summary Statistics
(Sample size = 4048)

Median error ratio	0.99836
Pseudo R^2	88.84%
COD	13.89
Regressivity	1.03459

Model

$$\ln(exppsf_{it}) = g_0 + g_1 \ln(\widehat{expratio}_{it}) + g_2 \ln(\widehat{incpsf}_{it}) + \in_{it}$$

Dependent variable: ln(Expense Per Square Foot)

Variable	Parameter estimate	Standard error	*t*-value	Pr > |*t*|
Intercept	0.2416	0.029650	8.15	<.0001
Pred_expratio	1.2286	0.015240	80.62	<.0001
Pred_incpsf	1.0092	0.008057	125.26	<.0001

Figure B3. Panel expense regression—Time-varying coefficients

Summary Statistics
(Sample size = 4048)

Median error ratio	0.9926
Pseudo R^2	90.80%
COD	13.01
Regressivity	1.0294

Model

$$\ln(exppsf_{it}) = g_{0t} + g_{1t}\ln(\widehat{expratio}_{it}) + g_{2t}\ln(\widehat{incpsf}_{it}) + \in_{it}$$

$$g_{0t} = f_0 + \gamma_{0t}$$
$$g_{1t} = f_1 + \gamma_{1t}$$
$$g_{2t} = f_2 + \gamma_{2t}$$

Dependent variable: ln(Expense Per Square Foot)

Variable	Parameter estimate	Standard error	*t*-value	Pr > |*t*|
Intercept	0.2529	0.14850	1.70	0.1872
Pred_expratio	1.2341	0.10060	12.27	0.0012
Pred_incpsf	1.0075	0.02141	47.05	<.0001

Tests of Time-varying Coefficients (*t*-value tests of $\gamma_{0t}, \gamma_{1t}, \gamma_{2t}$)

Effect	Period	Estimate	Std Err	*t*-value	Pr > |*t*|
Intercept	2010	−0.207800	0.15240	−1.36	0.1729
Pred_expratio	2010	0.006422	0.10250	0.06	0.9500
Pred_incpsf	2010	0.040700	0.02332	1.74	0.0811
Intercept	2011	0.424700	0.15270	2.78	0.0054
Pred_expratio	2011	0.247400	0.10250	2.41	0.0158
Pred_incpsf	2011	−0.006560	0.02336	−0.28	0.7789
Intercept	2012	−0.099110	0.15290	−0.65	0.5168
Pred_expratio	2012	−0.019030	0.10250	−0.19	0.8527
Pred_incpsf	2012	0.014360	0.02338	0.61	0.5392
Intercept	2013	−0.117800	0.15320	−0.77	0.4418
Pred_expratio	2013	−0.234800	0.10240	−2.29	0.0220
Pred_incpsf	2013	−0.048490	0.02345	−2.07	0.0387

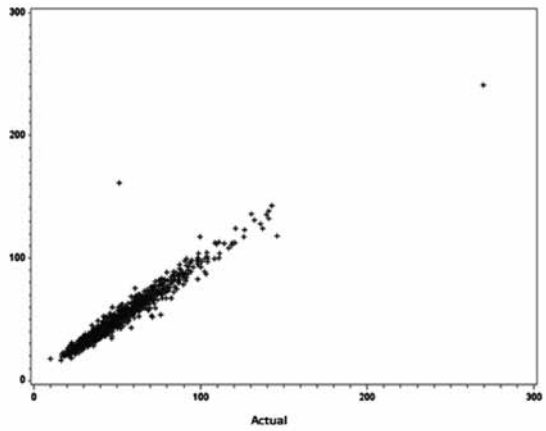
Figure B4. Predicted and actual income and expense for 2013 filers

Variable	Number	Minimum	25th percentile	50th percentile	75th percentile	Maximum	Mean
Actual Income PSF	1012	$9.93	$35.79	$45.84	$60.94	$269.84	$50.85
Predicted Income PSF	1012	$16.60	$36.27	$45.53	$60.08	$241.20	$50.73
Actual Expense PSF	1012	$2.82	$10.33	$13.54	$17.25	$57.78	$14.19
Predicted Expense PSF	1012	$4.34	$10.37	$13.33	$17.00	$46.93	$14.03
Pred NOI PSF	1012	$10.44	$25.03	$32.30	$44.13	$194.26	$36.70
Actual NOI PSF	1012	$5.83	$24.44	$31.85	$44.00	$212.06	$36.66

Income PSF 2013

Summary Statistics

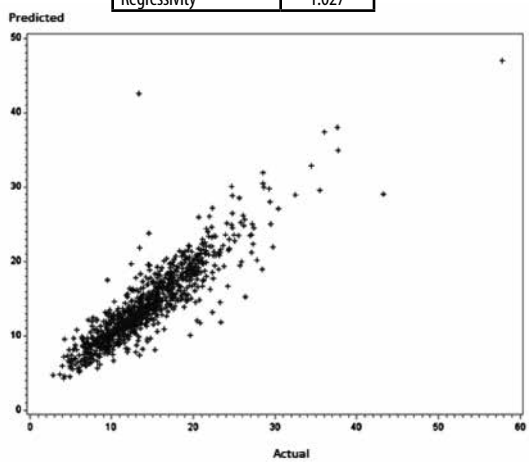
Median error ratio	1.002
Pseudo R^2	96.90%
COD	6.58
Regressivity	1.011



Expense PSF 2013

Summary Statistics

Median error ratio	0.991
Pseudo R^2	89.54%
COD	12.70
Regressivity	1.027



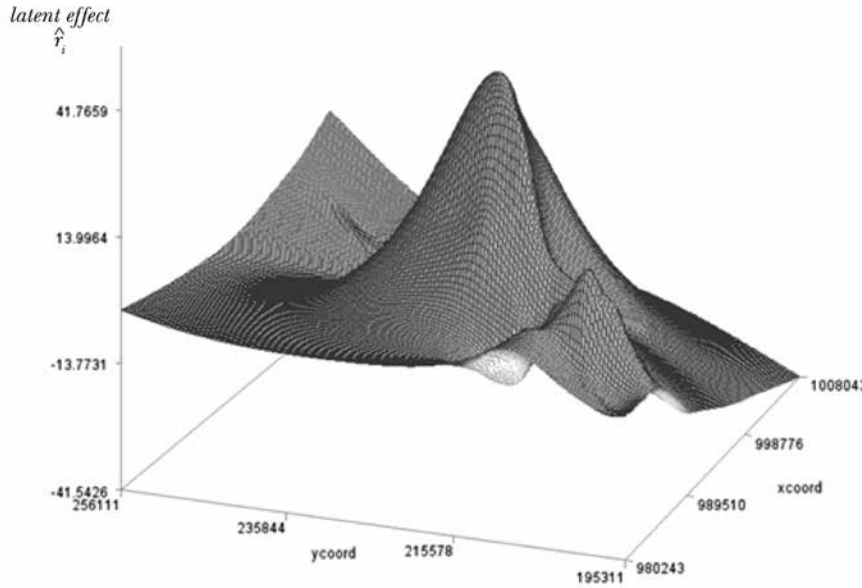
Appenidix C. Results of Hold-out Sample Test of Nonfiler Model Methodology

Figure C1. Latent effect models—Hold-out sample estimates

Income PSF
Summary Statistics
(Sample size = 254)

Median error ratio	1.0117
Pseudo R^2	66.00%
COD	26.23
Regressivity	1.1033

Spatial Distribution of Latent Effect Estimate



Actual and Predicted Income PSF

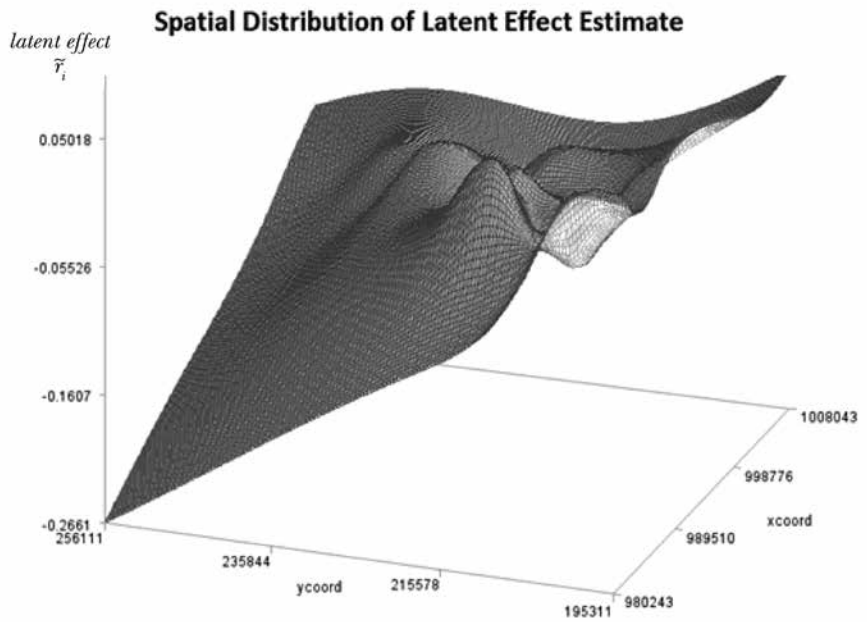
Variable	Number	5th percentile	25th percentile	50th percentile	75th percentile	95th percentile	Mean
Actual Income PSF	254	$25.89	$35.54	$47.20	$59.81	$89.20	$51.05
Predicted Income PSF	254	$32.42	$40.59	$47.17	$56.62	$76.70	$50.13

(figure continued on next page)

Figure C1. Latent effect models—Hold-out sample estimates (*continued*)

Expense Ratio
Summary Statistics
(Sample size = 254)

Median error ratio	0.9941
Pseudo R^2	52.02%
COD	17.67
Regressivity	1.0511



Actual and Predicted Expense Ratios

Variable	Number	5th percentile	25th percentile	50th percentile	75th percentile	95th percentile	Mean
Actual Expense Ratios	254	17.30%	25.71%	29.38%	33.00%	38.29%	28.82%
Predicted Expense Ratios	254	24.10%	27.18%	29.28%	31.02%	33.90%	29.14%

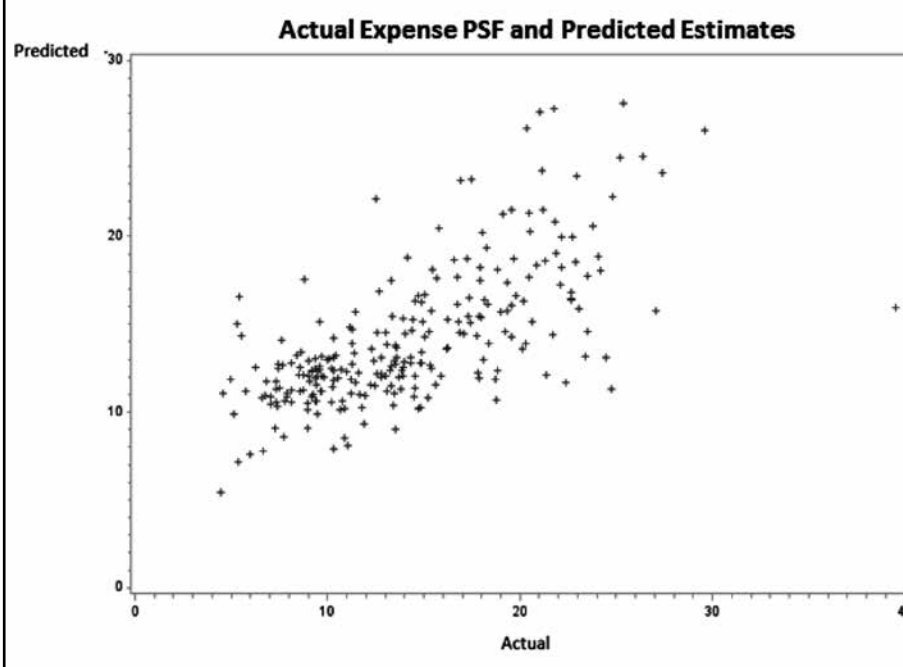
Figure C1. Latent effect models—Hold-out sample estimates (*continued*)

Expense PSF

Summary Statistics

(Sample size = 254)

Median error ratio	1.0205
Pseudo R^2	67.09%
COD	25.36
Regressivity	1.0979



Actual and Predicted Expense PSF

Variable	Number	5th percentile	25th percentile	50th percentile	75th percentile	95th percentile	Mean
Actual Expense PSF	254	$6.63	$9.59	$13.54	$18.08	$23.78	$14.25
Predicted Expense PSF	254	$9.90	$11.57	$13.01	$16.13	$22.23	$14.18

Appendix C (continued)

Table C1. Hold-out sample estimates—Model comparison with and without latent effects

(Hold-out sample size = 254)

Income		
	Without latent effects	**With latent effects**
Median error ratio	1.0136	1.0117
Pseudo R^2	53.46%	66.00%
COD	27.70	26.23
Regressivity	1.1109	1.1033
Expense Ratio		
	Without latent effects	**With latent effects**
Median error ratio	0.9975	0.9941
Pseudo R^2	26.53%	52.02%
COD	20.08	17.67
Regressivity	1.0568	1.0511
Expense		
	Without latent effects	**With latent effects**
Median error ratio	0.9889	1.0205
Pseudo R^2	55.21%	67.09%
COD	31.91	25.36
Regressivity	1.1390	1.0979