製表技術参考資料８

ISSN 1348－7345

サービス業基本調査における経理項目の補定法

N S T A C

Working Paper No.8

平成 20 年 3 月

独立行政法人 統計センター

製表技術参考資料は、独立行政法人 統計センターの職員がその業務に関連して行った製表技術に関する研究の結果を紹介するためのものである。

ただし、本資料の内容は執筆者の個人的見解を示すものであり、機関の見解を示すものではない。

目　次

サービス業基本調査における経理項目の補定法

村田　磨理子*，畠山　昌子**，磯部　祥子**，亀本　薫**

要　旨

本稿は、サービス業基本調査における給与支給総額、収入額等の経理項目を対象として、欠測値の補定法に関する検討を行った結果を取りまとめたものである。

検討では、平成 16 年サービス業基本調査を対象とし、経理項目に対する回帰補定の改善の可能性を調べた。現行の補定法では、例えば、給与支給総額は、１雇用者当たり給与支給総額に当該事業所の雇用者数を乗じた値で補定する。つまり、雇用者数を説明変数とし、それに一定の比率（1 雇用者当たり給与支給総額）を乗じる形であることから、定数項のない回帰モデルに相当する。ここで、1 雇用者当たり給与支給総額は、事業所の産業、規模、地域などによる層ごとに算出される。

まず予備的な分析において、定数項の適用や変数変換の必要性が示唆された。さらに、回帰モデルを構築するためには、完全データのサイズが小さい層が多いという問題があった。本研究では、線形回帰モデルを拡張した対数線形回帰モデルと１人当たり回帰モデルの当てはめを行い、回帰の予測残差に基づくモデル比較の方法を提案した。予測残差による評価や外れ値の検出に一定の効果があることを示したが、さらなる検討や業種業態などの背景領域の知識が必要であると思われる。また、層の構築については、層化されていない単回帰モデルの残差を多段層別分析によって分類する方法と、現行の補定法における層化変数を説明変数に入れた重回帰・数量化Ｉ類モデルよって層化因子の効果を確認する方法を提案した。どちらも回帰モデルの構築と統合的な方法であるが、２つの方法で得られた層化が一致しない場合がある。業種業態などの背景領域の情報を取り入れるといった専門的な判断が必要と考えられる。

* 統計センター研究センター非常勤職員（シンフォニカ主任研究員）
** 統計センター研究センター（E-mail: research@nstac.go.jp）

サービス業基本調査における経理項目の補定法

村田　磨理子，畠山　昌子，磯部　祥子，亀本　薫

１．はじめに

（１）研究の目的

研究センターでは、主要な調査項目でありながら非回答が比較的多く、補定の精度の影響が特に大きい調査項目について、補定法の改善の研究を進めている。平成 15、16 年度は全国消費実態調査の年収を対象に研究を行い、Dissimilarity matrix 法や Predictive mean matching 法など最近隣補定法あるいは最近隣補定法と回帰補定法を組合せた方法を用いることにより補定精度の向上が期待できることを示した。

しかし、事業所や企業を対象とする調査（以下、事業所・企業調査という）は、世帯の調査とは特性が異なるため、事業所・企業調査データの特性に合った新たな補定法を別途に研究する必要がある。そこで、平成 17 年度からサービス業基本調査の収入額、経費総額、給与支給総額を対象として、欠測値の補定法について研究を行うこととした。最初に、経理項目の補定の理論や適用事例の情報収集により研究方法の検討を行った。次に、平成 16 年サービス業基本調査の個別データを用いて、経理に関する項目の回帰補定を検討した。

本稿は、平成 16 年サービス業基本調査の個別データを用いて、現行の補定方法の改良を試みた結果をまとめたものである。

（２）平成 16 年サービス業基本調査の経理項目の補定について

平成 16 年サービス業基本調査は、経理に関する項目として、収入額、経費総額、給与支給総額、設備投資額を設けている。調査の設計上、設備投資額を除く経理３項目（以下、「経理項目」という）には不詳を設けていないが、地方でのデータチェックでは経理項目の欠測値を許容している。それらの欠測値は、統計センターにおいてデータ訂正及び補定処理等が行われる。

統計センターで実施した経理項目に関する主な補定処理は次のとおりである（統計センター事業企画課，2005）。

経理項目すべてに記入があり、設定された上限値あるいは下限値を超過しないデータ（以下、金額補定用基礎データという）を用い、産業、経営組織、従業者規模等による層ごとに次の方法で算出する。

- 給与支給総額は、金額補定用基礎データから算出した１雇用者当たり給与支給総額に当該事業所の雇用者数を乗じた値で補定する。
- 経費総額は、金額補定用基礎データから算出した給与支給総額に対する経費総額の割合に当該事業所の給与支給総額を乗じた値で補定する。
- 収入額は、単独事業所については、金額補定用基礎データから算出した経費総額に対する収入額の割合に当該事業所の経費総額を乗じた値で補定し、本所又は支所の事業所については、金額補定用データから算出した１従業者当たり収入額に当該事業所の従業者数を乗じた値で補定する。

ここで、層化は、産業分類、経営組織、本所・支所の別、従業者規模、地域のクロスによる。それぞれの層化変数は、以下の区分を持つ。

- 産業分類：産業小分類（サービス業基本調査結果表で表章する分類）
- 経営組織：「個人経営」、「会社」（株式会社、有限会社、合名会社・合資会社、相互会社、外国の会社）、「その他」（会社以外の法人、法人でない団体）
- 本所・支所の別：「単独事業所」、「本所・本社・本店」、「支所・支社・支店」
- 従業者規模：「0～4 人」、「5～29 人」、「30 人以上」
- 地域：「人口 30 万以上市」、「人口 30 万未満市」、「町村」

また、層内の金額補定用基礎データの事業所数が 4 以下の場合、層を地域、従業者規模の 0～29 人、本所・支所の別、経営組織、従業者規模（全階級）の順に合併することとしている。

２．予備的分析

（１）現行の方法の問題点

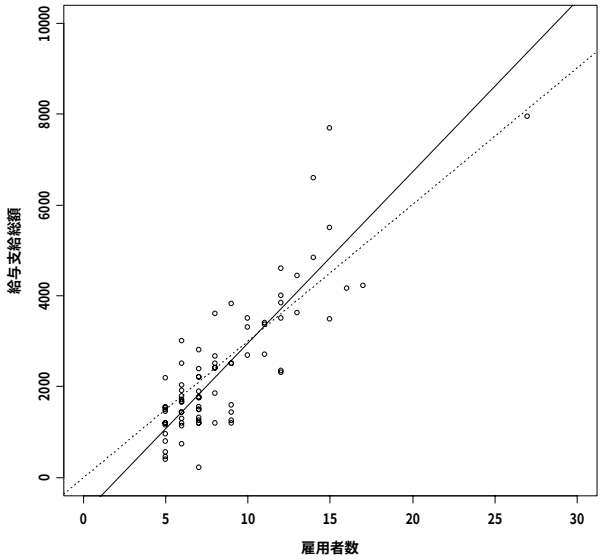
最初に、給与支給総額の補定に絞って、現行の方法の問題点を探る。給与支給総額の補定値の算出方法は、給与支給総額を被説明変数、雇用者数を説明変数とする回帰モデルを適用したものと見ることができる。説明変数に一定の比率（1 雇用者当たり給与支給総額）を乗じる形であることから、特に、定数項のない回帰モデル（以下、比推定型回帰モデルという）に相当する。なお、回帰モデルは前述の層ごとに構築することになる。

ア　比推定型の妥当性

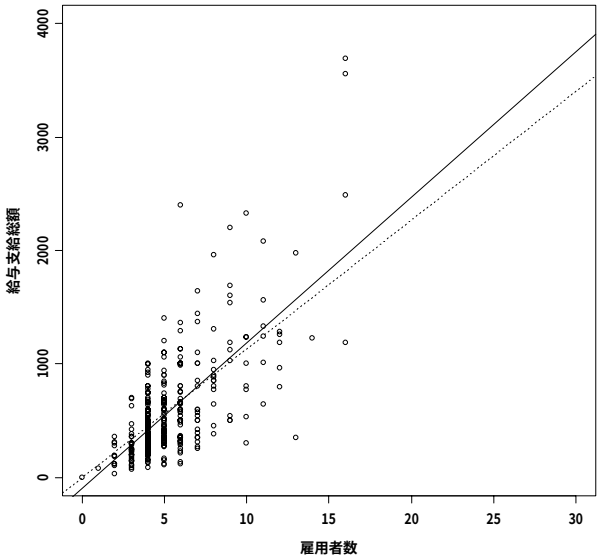
比推定型モデルは、雇用者数と給与支給総額の関係については、比例、つまり、原点を通る直線を仮定している。最初に、すべての層について、雇用者数と給与支給総額の散布図を作成し、２つの変数の関係を観察した。その中で、比推定型モデルの仮定と反して、原点を通らないと考えられる例（図１－１）や、直線の関係ではなく、両対数を取って直線に近くなると考えられる例（図１－２）を示す。

図１－１　原点を通らないと考えられる例

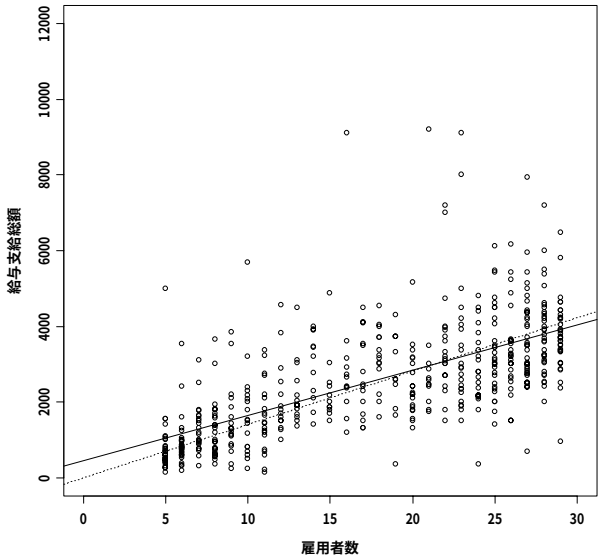
① 823 美容業



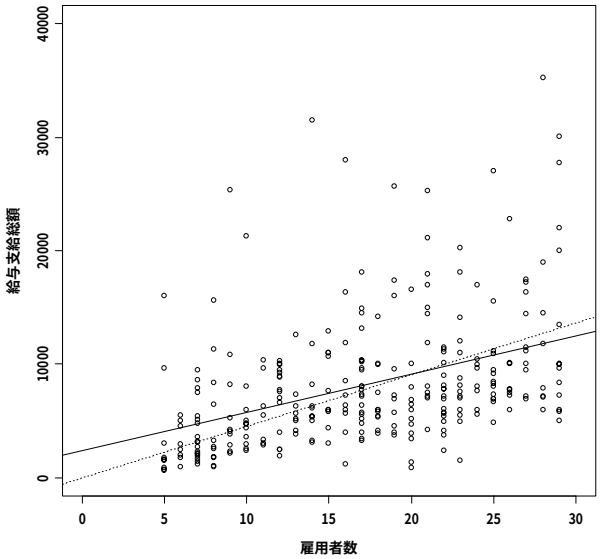
② 70A 一般食堂



③ 70A 一般食堂



④ 84J パチンコホール

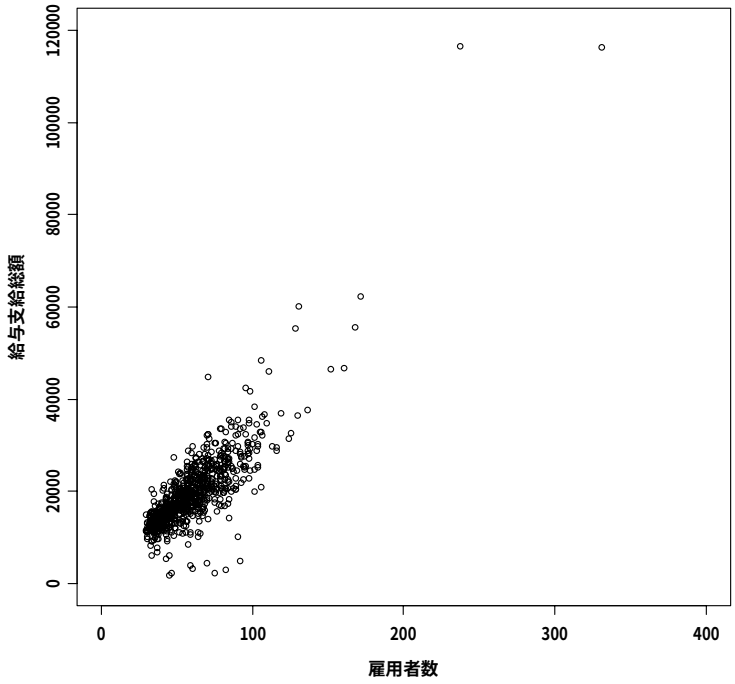


（原点を通る直線を点線で示している。）

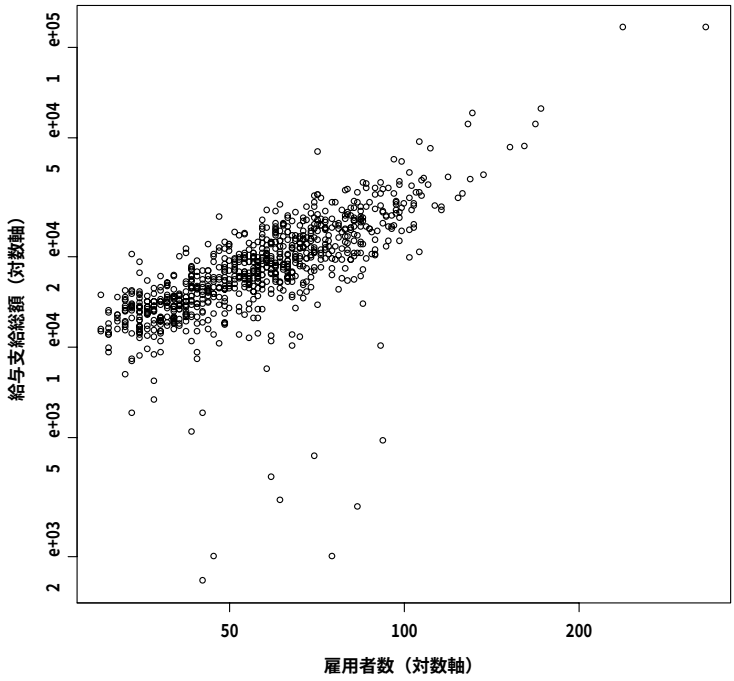
①「823 美容業、会社、支所・支社・支店、5～29 人、人口 30 万未満市」
②「70A 一般食堂、個人経営、単独事業所、5～29 人、人口 30 万未満市」
③「70A 一般食堂、会社、支所・支社・支店、5～29 人、県庁所在市・人口 30 万以上市」
④「84J パチンコホール、会社、本所・本社・本店、5～29 人、県庁所在市・人口 30 万以上市」

図２－２　両対数を取って直線に近くなる例

⑤－１　75C 特別養護老人ホーム　原数値



⑤－２　75C 特別養護老人ホーム　両対数



⑤「75C 特別養護老人ホーム、その他、単独事業所、30 人以上、町村」の場合。原数値の散布図（⑤－１）は、右上にむかって曲線を描く様子が観察されるが、両対数の散布図では（⑤－２）は、直線により近くなる。

イ　安定した推定のためのサンプルサイズ

この研究では、データ項目の中の「補定対象事業所（補定の種類）」のフラグを使って、経理項目が記入されているかどうかを識別した。経理項目がすべて記入されている事業所のデータを、以下では、完全データという。回帰モデルの構築は、完全データのみを使うため、サンプルサイズが小さいと当てはめが不安定になる。現行の層化における各層の完全データのサンプルサイズを表２－１にまとめる。ただし、ここでは、前述の層の合併を行っていない。

表２－１から、全産業は 5,811 層に分割され、完全データの存在しない層が 284 層、完全データのサンプルサイズが 1～4 の層が 2,230 層であることが分かる。一方で、完全データのサンプルサイズが十分大きい 100 以上の層は、538 層に過ぎず、サンプルサイズの小さい層が多いことが分かる。

表２－１　完全データのサンプルサイズ階級別層数

完全データのサンプルサイズ階級	全産業計	うち分析対象の小分類									
		82A 普通洗濯業	823 美容業	84J パチンコホール	75H 訪問介護事業	70A 一般食堂	882 産業用機械器具賃貸業	906 警備業	90A 労働者派遣業	691 不動産賃貸業（貸家業，貸間業を除く）	75C 特別養護老人ホーム
計	5,811	67	49	48	54	69	50	50	46	52	20
0	284	0	1	2	0	1	2	2	2	1	0
1～4	2,230	22	10	6	21	18	17	20	15	22	6
5～9	909	9	9	6	9	11	1	5	7	8	2
10～19	769	5	8	5	8	10	4	7	9	5	2
20～29	367	5	3	2	2	2	5	1	1	4	1
30～39	218	3	3	5	5	3	3	3	2	1	0
40～49	162	7	2	2	3	5	1	1	1	1	0
50～99	334	8	8	5	4	2	5	1	3	4	0
100～199	238	3	2	5	2	6	6	4	4	2	4
200 以上	300	5	3	10	0	11	6	6	2	4	5

（2）1 変量モデルの適用

給与支給総額を被説明変数、雇用者数を説明変数とする 1 変量のモデルを検討する。ここでは、定数項の有無、変数変換などが異なる複数のモデルを比較した。具体的には、従業者規模 0～4 人のケースは、①定数項の有無を比較、②雇用者数及び給与支給総額が 0 でないデータに限定して定数項の有無を比較する。また、従業者規模 5 人以上のケースは、①定数項の有無の比較、②原数値と対数変換を比較する。

この予備的分析では、いくつかの代表的な層に限定して比較を行った。まず、経費総額に占める給与支給総額の割合、サービスの提供対象による分類などを参考にして、典型的な産業を選定し（表２－２）、次に、それぞれの産業の中で、完全データのサンプルサイズの大きい層を１つずつ選定して、検討対象を絞った（表２－３）。選定した各層における散布図、推定結果の残差の統計量等を、次のア～コに示す。

表２－２　選定した典型的な産業の特徴

産業小分類		特徴
82A	普通洗濯業	対個人，生活関連サービスの典型，小規模事業所が多い，1雇用者当たり給与が低い
823	美容業	対個人，生活関連サービスの典型，小規模事業所が多い，1雇用者当たり給与が高い
84J	パチンコホール	対個人，給与支給総額割合が低い
75H	訪問介護事業	対個人（福祉），給与支給総額割合が極めて高い
70A	一般食堂	対個人，直接消費財を供給，H16 調査の新規対象産業
882	産業用機械器具賃貸業	対事業所，物的
906	警備業	対事業所，給与支給総額割合が高い，物的･人的･システム的の混合
90A	労働者派遣業	対事業所，給与支給総額割合が高い，人的（システム的な側面も持つ）
691	不動産賃貸業（貸家業，貸間業を除く）	対事業所，H16 調査の新規対象産業
75C	特別養護老人ホーム	経営組織が「その他」（＝いわゆる民間非営利）が多い

表２－３　分析対象の層

産業小分類		経営組織	本所･支所の別	従業者規模	地域	完全データのサンプルサイズ
82A	普通洗濯業	個人経営	単独事業所	0～4 人	人口 30 万未満市	1,038
823	美容業	個人経営	単独事業所	0～4 人	人口 30 万未満市	2,071
84J	パチンコホール	会社	単独事業所	5～29 人	人口 30 万未満市	531
75H	訪問介護事業	会社	単独事業所	30 人以上	県庁所在市･人口 30 万以上市	150

70A	一般食堂	会社	支所･支社･支店	30 人以上	県庁所在市･人口 30 万以上市	982
882	産業用機械器具賃貸業	会社	支所･支社･支店	5～29 人	人口 30 万未満市	393
906	警備業	会社	単独事業所	30 人以上	県庁所在市･人口 30 万以上市	538
90A	労働者派遣業	会社	単独事業所	30 人以上	県庁所在市･人口 30 万以上市	343
691	不動産賃貸業（貸家業，貸間業を除く）	会社	単独事業所	0～4 人	県庁所在市･人口 30 万以上市	407
75C	特別養護老人ホーム	その他	単独事業所	30 人以上	町村	822

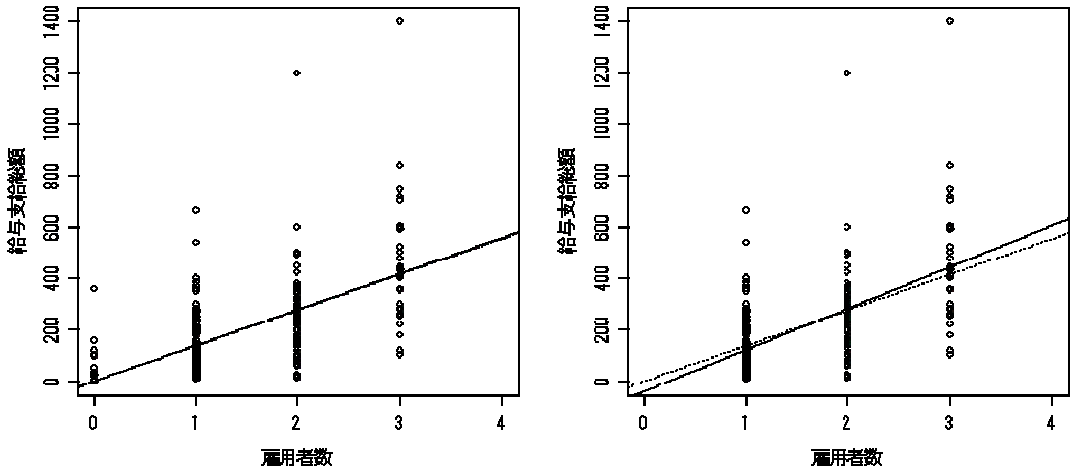
ア　「82A 普通洗濯業、個人経営、単独事業所、0～4 人、人口 30 万未満市」における雇用者数と給与支給総額

①定数項の有無を比較（左のグラフ）

定数項は 0 に近い値となり、有意ではない。また、雇用者数が大きくなるにつれて、回帰直線の周りのバラツキが大きくなる傾向が見られる。このことから、回帰モデルの誤差項の分散が一定であるとする仮定は、適切ではない可能性がある。

②雇用者数及び給与支給総額が 0 でないデータに限定して、定数項の有無を比較（右のグラフ）

定数項ありのモデルの方が AIC の値が少し小さいが、定数項は 5%水準では有意でない。



n=1,038

対象（注）	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
0 を含む	なし	雇用者数	139.26	2.91	47.78	0.000	81.34	0.687	12080.27
	あり	定数項	-1.09	2.88	-0.38	0.706	81.37	0.631	12082.13
		雇用者数	139.86	3.32	42.10	0.000			
0 を含まない	なし	雇用者数	139.26	5.46	25.52	0.000	152.31	0.693	3715.22
	あり	定数項	-40.21	21.55	-1.87	0.063	151.65	0.345	3713.73
		雇用者数	161.50	13.10	12.33	0.000			

（注）雇用者数及び給与支給総額が 0 を含むか否かを示す。以下、同様。

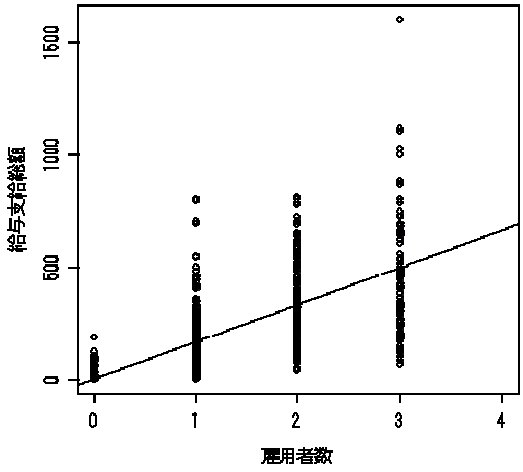
イ　「823 美容業、個人経営、単独事業所、0～4 人、人口 30 万未満市」における雇用者数と給与支給総額

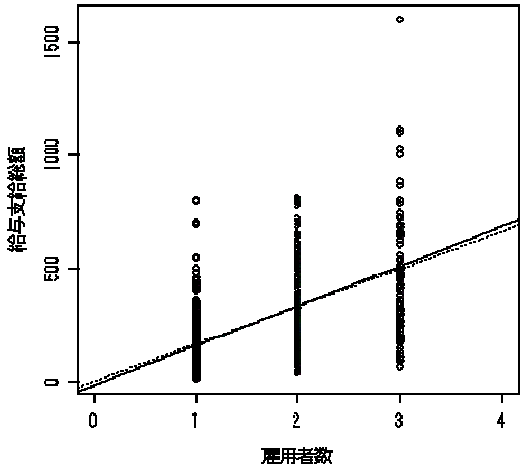
①定数項の有無を比較（左のグラフ）

定数項は 0 に近い値となり、有意ではない。

②雇用者数及び給与支給総額が 0 でないデータに限定して、定数項の有無を比較（右のグラフ）

定数項なしのモデルの方が AIC の値がわずかに小さい。





n=2,071

対象	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
0 を含む	なし	雇用者数	165.82	2.18	76.09	0.000	98.19	0.737	24879.27
	あり	定数項	-0.58	2.58	-0.23	0.821	98.21	0.663	24881.22
		雇用者数	166.14	2.60	63.84	0.000			
0 を含まない	なし	雇用者数	165.90	3.61	45.93	0.000	162.71	0.738	9743.28
	あり	定数項	-18.70	14.22	-1.32	0.189	162.63	0.358	9743.55
		雇用者数	176.21	8.63	20.42	0.000			

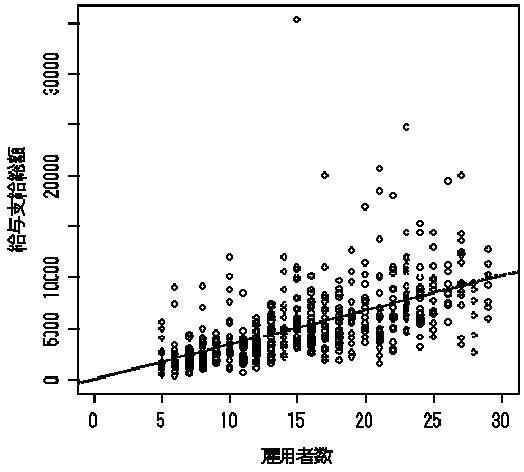
ウ　「84J パチンコホール、会社、単独事業所、5～29 人、人口 30 万未満市」における雇用者数と給与支給総額

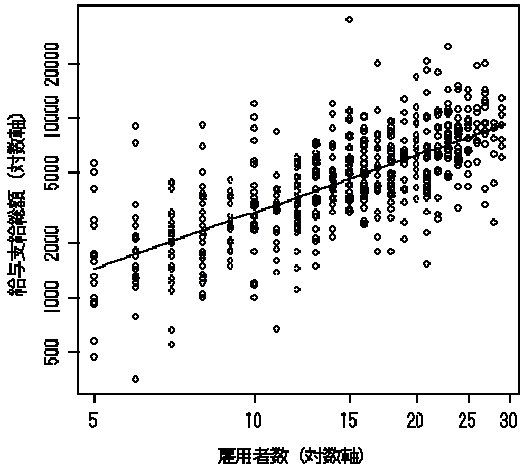
①定数項の有無を比較（左のグラフ）

定数項は有意ではない。残差の標準偏差は、どちらもかなり大きい。

②雇用者数及び給与支給総額を対数変換（右のグラフ）

雇用者数（対数）の係数は、かなり１に近い（つまり、ほぼ線形モデルと同じ）。





n=531

変数変換	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
変換なし	なし	雇用者数	341.96	7.75	44.11	0.000	3023.92	0.786	10021.11
	あり	定数項	74.26	352.54	0.21	0.833	3026.65	0.331	10023.06
		雇用者数	337.88	20.82	16.23	0.000			
対数変換		定数項	5.60	0.13	44.40	0.000	0.49	0.485	749.01
		雇用者数(対数)	1.04	0.05	22.35	0.000			

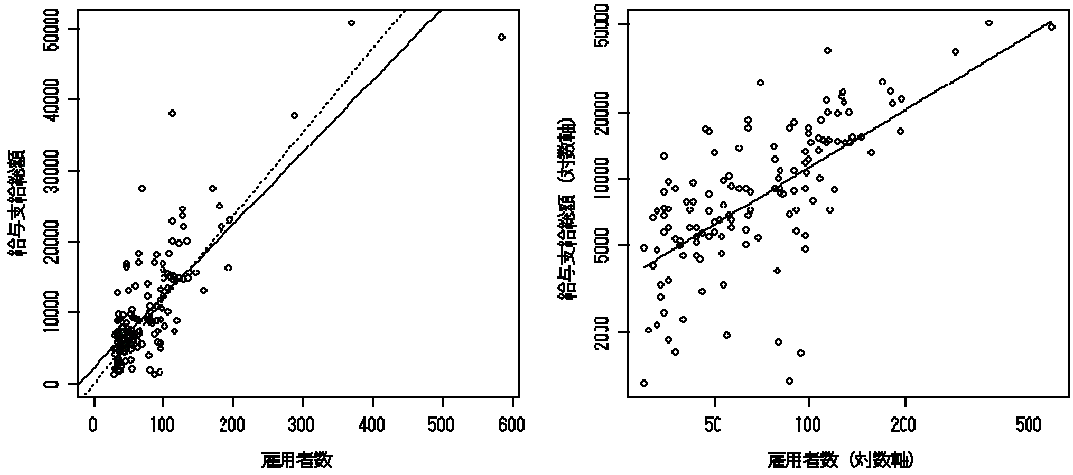
エ　「75H 訪問介護事業、会社、単独事業所、30 人以上、県庁所在市・人口 30 万以上市」における雇用者数と給与支給総額

①定数項の有無を比較（左のグラフ）

定数項は有意で、かなり大きい値となる。定数項のあるモデルのほうが、残差の標準偏差が小さくなり、AIC も小さい。

②雇用者数及び給与支給総額を対数変換（右のグラフ）

雇用者数（対数）の係数は、１より小さい。



変数変換	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
変換なし	なし	雇用者数	118.32	4.02	29.45	0.000	5052.27	0.852	2986.96
	あり	定数項	2142.18	654.75	3.27	0.001	4895.37	0.630	2978.48
		雇用者数	101.79	6.38	15.96	0.000			
対数変換		定数項	5.36	0.34	15.72	0.000	0.53	0.438	238.19
		雇用者数(対数)	0.86	0.08	10.82	0.000			

n=150

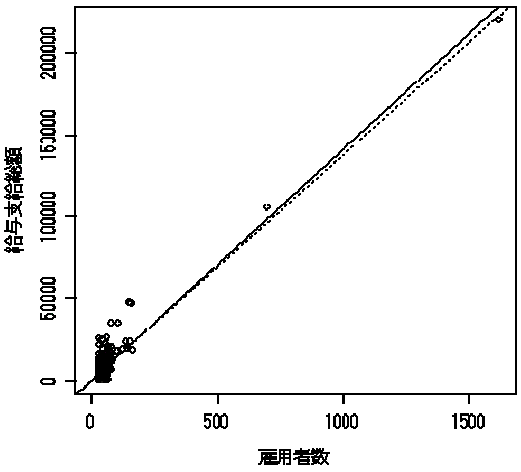
オ　「70A 一般食堂、会社、支所･支社･支店、30 人以上、県庁所在市・人口 30 万以上市」における雇用者数と給与支給総額

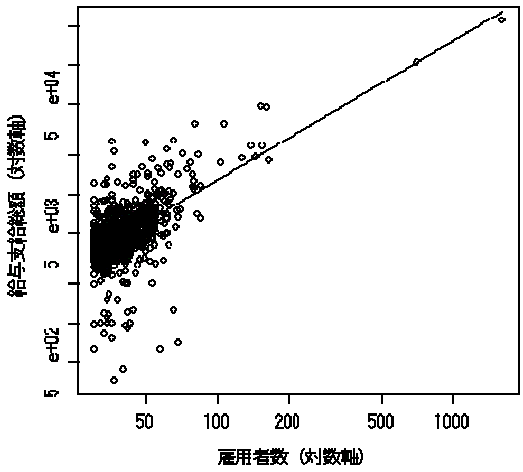
①定数項の有無を比較（左のグラフ）

定数項は有意で、マイナスの値となる。残差の標準偏差はあまり変わらないが、AIC は定数項のあるモデルの方が小さい。

②雇用者数及び給与支給総額を対数変換（右のグラフ）

雇用者数（対数）の係数は、かなり 1 に近い。

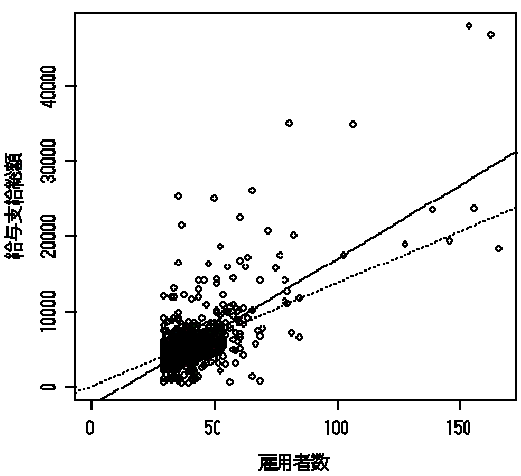


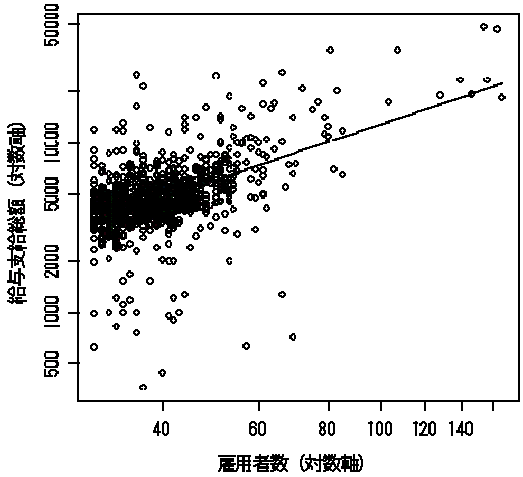


n=982

変数変換	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
変換なし	なし	雇用者数	137.80	1.29	106.50	0.000	2873.46	0.920	18429.66
	あり	定数項	-435.32	115.27	-3.78	0.000	2854.23	0.885	18417.47
		雇用者数	141.57	1.63	87.03	0.000			
対数変換		定数項	4.48	0.18	25.52	0.000	0.41	0.345	1038.26
		雇用者数(対数)	1.08	0.05	22.77	0.000			

次に、点の集中している場所を観察しやすくするため、雇用者数と給与支給総額がともに大きい２事業所を除外して、比較する。定数項のあるモデルは、定数項の推定値が大幅に下がり、雇用者数の係数が大きくなる。





n=982

変数変換	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
変換なし	なし	雇用者数	137.12	2.11	64.88	0.000	2859.46	0.811	18382.56
	あり	定数項	-2691.16	286.93	-9.38	0.000	2740.34	0.472	18300.16
		雇用者数	196.42	6.64	29.59	0.000			
対数変換		定数項	4.44	0.21	21.30	0.000	0.41	0.276	1037.96
		雇用者数（対数）	1.09	0.06	19.32	0.000			

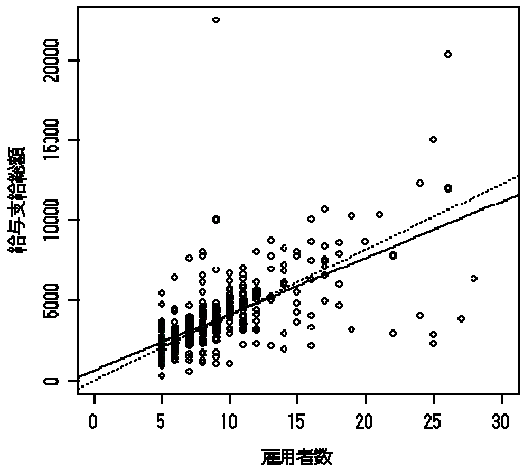
カ　「882 産業用機械器具賃貸業、会社、支所･支社･支店、5～29 人、人口 30 万未満市」における雇用者数と給与支給総額

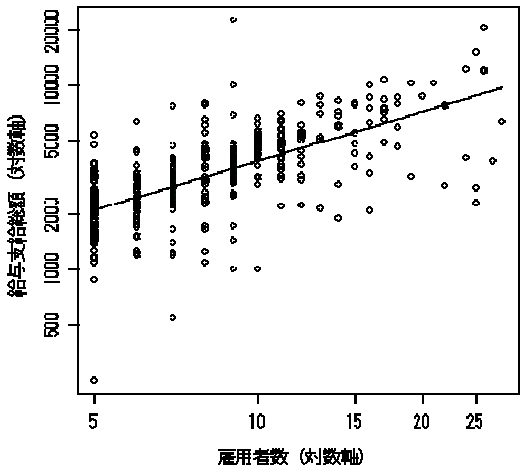
①定数項の有無を比較（左のグラフ）

定数項は有意で、プラスの値となる。残差の標準偏差はあまり変わらないが、AIC は定数項のあるモデルの方が小さい。

②雇用者数及び給与支給総額を対数変換（右のグラフ）

雇用者数（対数）の係数は、１より小さい。





n=393

変数変換	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
変換なし	なし	雇用者数	408.74	9.61	42.53	0.000	1908.63	0.821	7055.84
	あり	定数項	626.32	215.93	2.90	0.004	1890.83	0.405	7049.47
		雇用者数	352.65	21.56	16.36	0.000			
対数変換		定数項	6.22	0.11	57.20	0.000	0.41	0.436	424.20
		雇用者数（対数）	0.89	0.05	17.45	0.000			

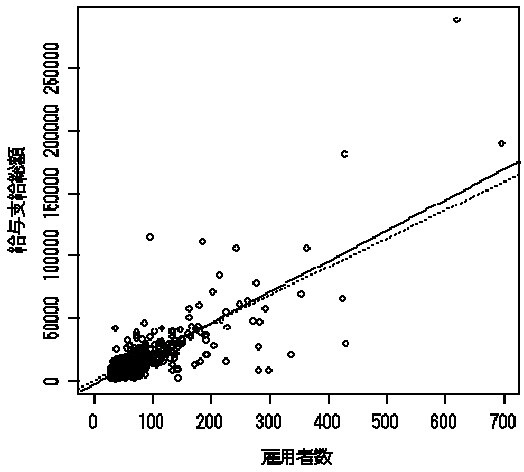
キ　「906 警備業、会社、単独事業所、30 人以上、県庁所在市・人口 30 万以上市」における雇用者数と給与支給総額

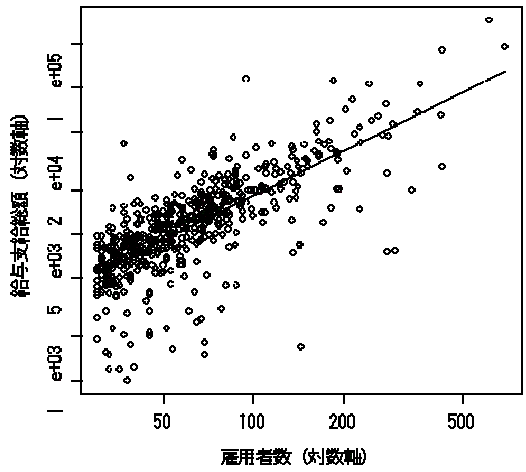
①定数項の有無を比較（左のグラフ）

定数項は有意で、マイナスの値となる。残差の標準偏差はあまり変わらないが、AIC は定数項のあるモデルの方が小さい。

②雇用者数及び給与支給総額を対数変換（右のグラフ）

雇用者数（対数）の係数は、ほぼ１である。





n=538

変数変換	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
変換なし	なし	雇用者数	226.79	5.36	42.30	0.000	12865.88	0.769	11711.25
	あり	定数項	-2634.83	835.02	-3.16	0.002	12759.91	0.633	11703.35
		雇用者数	245.95	8.07	30.47	0.000			
対数変換		定数項	5.20	0.17	31.05	0.000	0.53	0.541	842.28
		雇用者数(対数)	1.01	0.04	25.18	0.000			

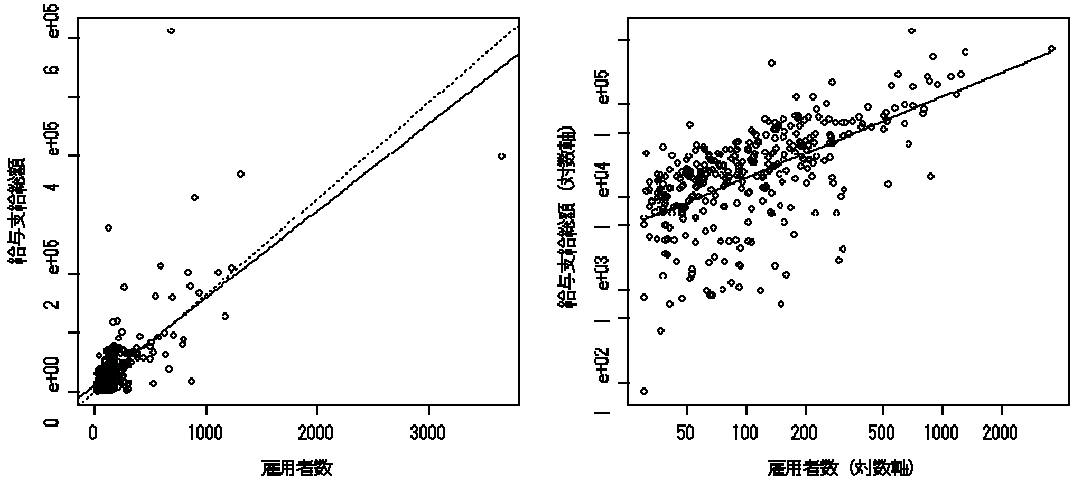
ク　「90A 労働者派遣業、会社、単独事業所、30 人以上、県庁所在市・人口 30 万以上市」における雇用者数と給与支給総額

①定数項の有無を比較（左のグラフ）

定数項は有意で、プラスの値となる。残差の標準偏差と AIC は、ともに定数項のあるモデルの方が小さい。

②雇用者数及び給与支給総額を対数変換（右のグラフ）

雇用者数（対数）の係数は、１より小さい。



n=343

変数変換	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
変換なし	なし	雇用者数	164.26	6.96	23.61	0.000	41639.76	0.619	8273.24
	あり	定数項	9584.86	2607.15	3.68	0.000	40898.14	0.497	8261.91
		雇用者数	148.49	8.07	18.41	0.000			
対数変換		定数項	5.69	0.29	19.29	0.000	1.01	0.365	982.39
		雇用者数(対数)	0.87	0.06	14.05	0.000			

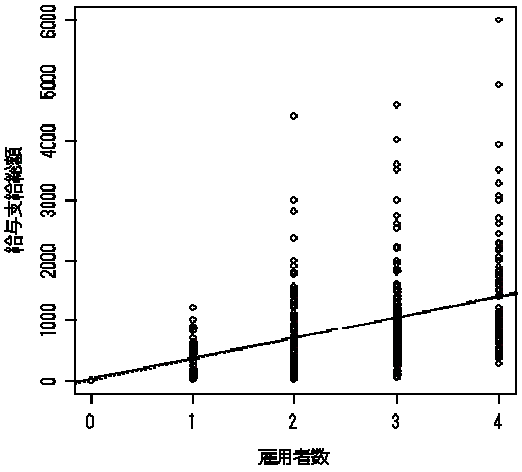
ケ　「691 不動産賃貸業（貸家業，貸間業を除く）、会社、単独事業所、0～4 人、県庁所在市・人口 30 万以上市」における雇用者数と給与支給総額

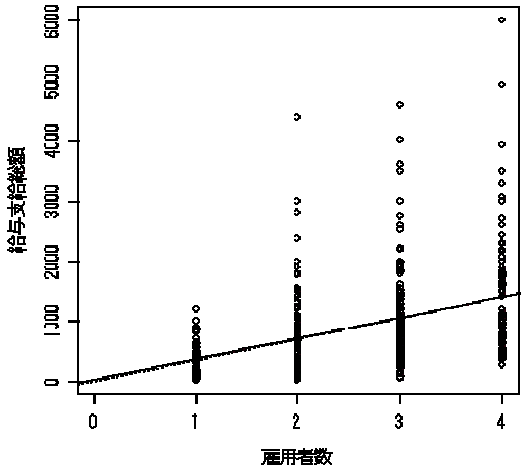
①定数項の有無を比較（左のグラフ）

定数項は 0 に近い値となり、有意ではない。

②雇用者数及び給与支給総額が 0 でないデータに限定して、定数項の有無を比較（右のグラフ）

定数項なしのモデルの方が AIC の値がわずかに小さい。





n=407

対象	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
0 を含む	なし	雇用者数	353.16	13.39	26.37	0.000	742.83	0.630	6538.93
	あり	定数項	34.29	101.18	0.34	0.735	743.64	0.173	6540.82
		雇用者数	341.55	36.80	9.28	0.000			
0 を含まない	なし	雇用者数	353.16	13.46	26.24	0.000	746.52	0.630	6478.69
	あり	定数項	37.03	105.67	0.35	0.726	747.33	0.163	6480.57
		雇用者数	340.62	38.25	8.91	0.000			

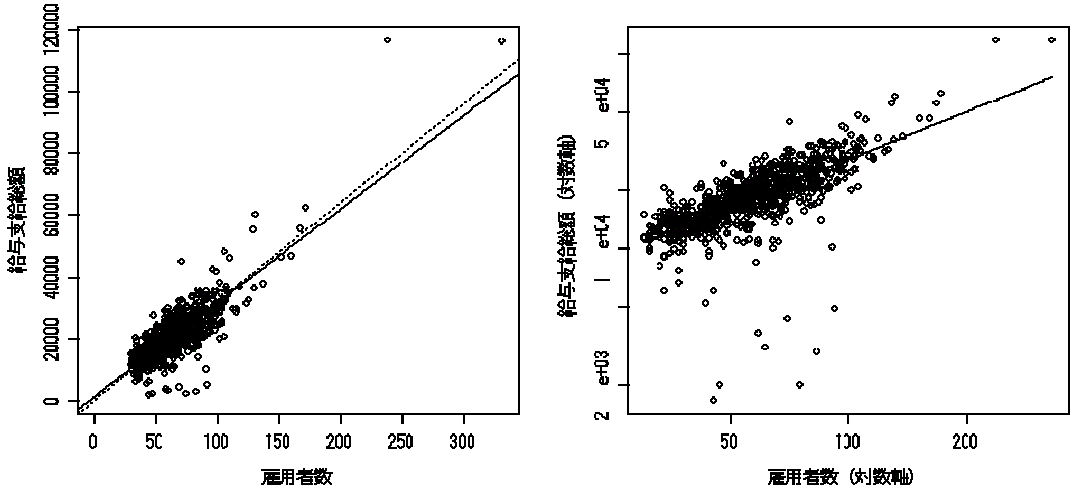
コ　「75C 特別養護老人ホーム、その他、単独事業所、30 人以上、町村」における雇用者数と給与支給総額

①定数項の有無を比較（左のグラフ）

定数項は有意で、プラスの値となる。残差の標準偏差はあまり変わらないが、AIC は定数項のあるモデルの方が小さい。

②雇用者数及び給与支給総額を対数変換（右のグラフ）

雇用者数（対数）の係数は、１より小さい。



n=822

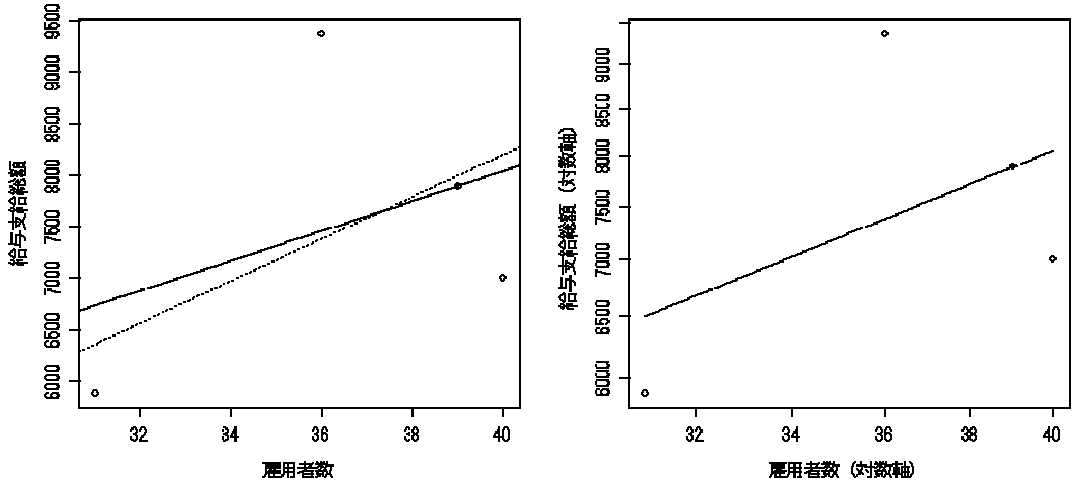
変数変換	定数項の有無	パラメータ	推定値	標準誤差	t値	Pr(>|t|)	残差の標準偏差	自由度修正決定係数	AIC
変換なし	なし	雇用者数	320.92	2.50	128.50	0.000	4609.73	0.953	16204.40
	あり	定数項	1202.08	434.19	2.77	0.006	4591.13	0.712	16198.75
		雇用者数	303.57	6.74	45.03	0.000			
対数変換		定数項	6.53	0.12	56.76	0.000	0.28	0.497	223.12
		雇用者数(対数)	0.81	0.03	28.48	0.000			

（３）層の合併

サンプルサイズが小さい場合、回帰モデルの誤差が相対的に大きくなるため、安定した推定が難しい。安定した推定を行うために、産業分類以外の地域、経営組織、従業者規模等の合併を一通り試みた。図２－３に、「82A 普通洗濯業、個人経営、単独事業所、30 人以上、人口 30 万未満市」の層（n=5）について、散布図と回帰モデルの推定結果の変化を示す。

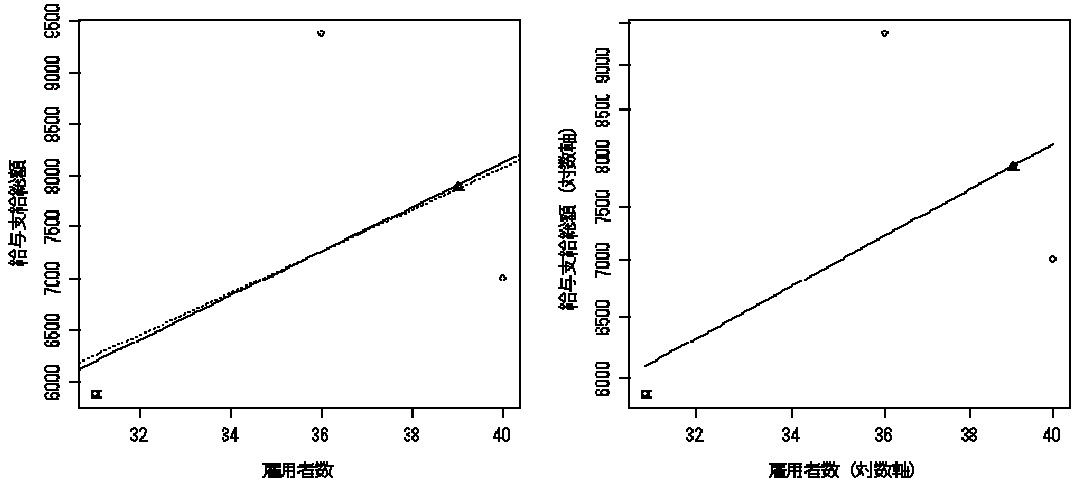
図２－３　層の合併による回帰当てはめの変化

ア　「82A 普通洗濯業、個人経営、単独事業所、30 人以上、人口 30 万未満市」（n=5）



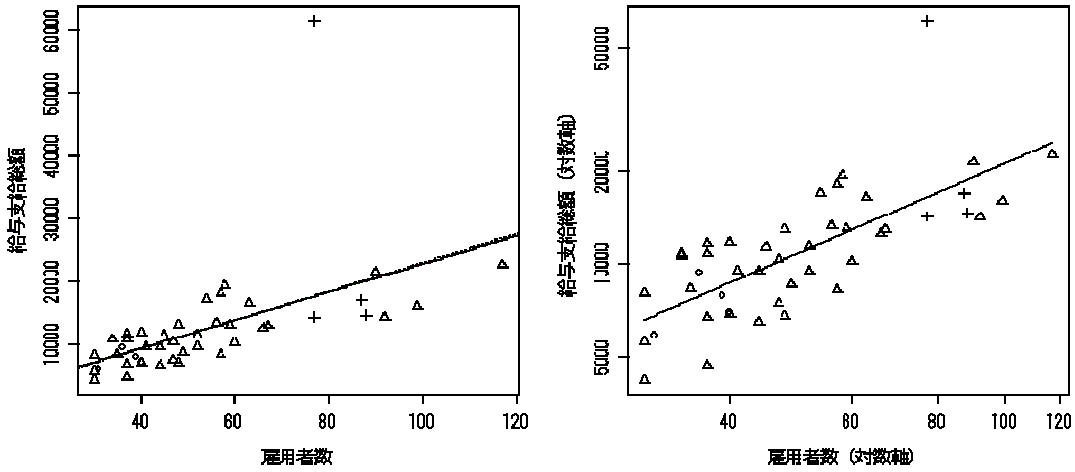
イ　地域を合併

「82A 普通洗濯業、個人経営、単独事業所、30 人以上、全地域」（n=8）



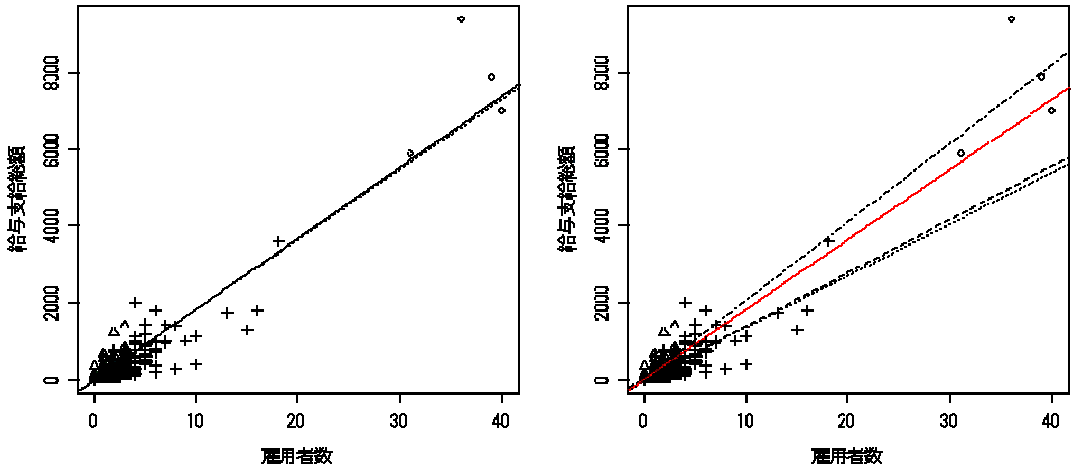
ウ　経営組織を合併

「82A 普通洗濯業、全経営組織、単独事業所、30 人以上、人口 30 万未満市」（n=46）



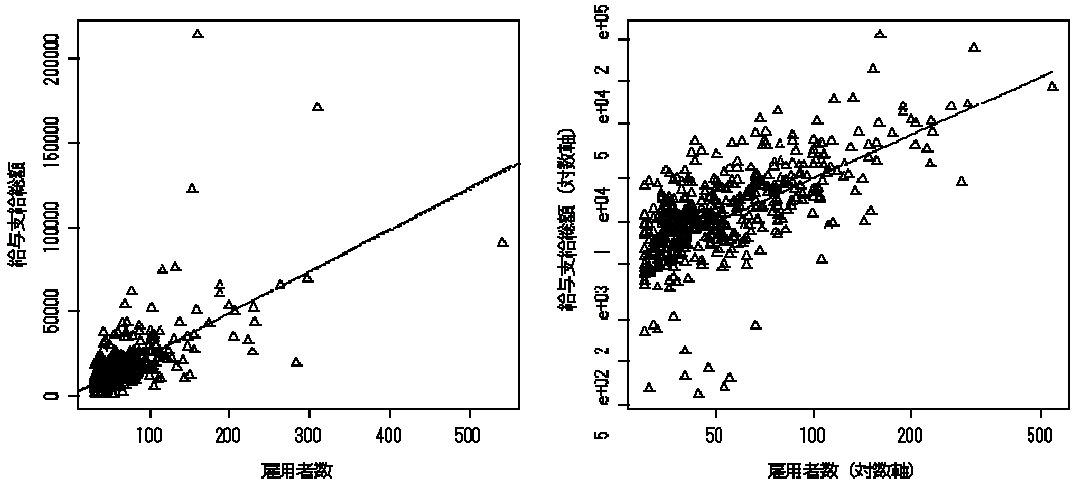
エ　従業者規模を合併

「82A 普通洗濯業、個人経営、単独事業所、全従業者規模、人口 30 万未満市」（n=1,108）



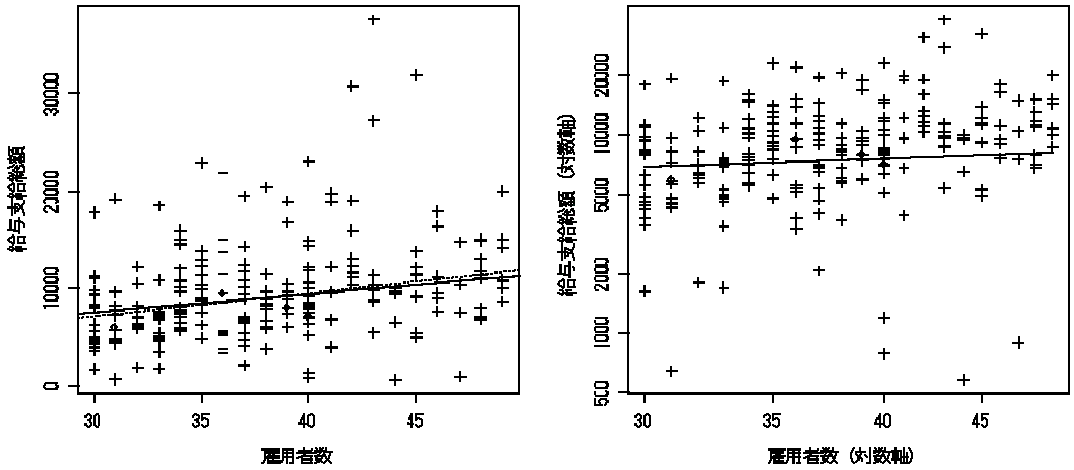
オ　従業者規模以外を合併

「82A 普通洗濯業、全経営組織、全事業所、30 人以上、全地域」（n=483）



カ　従業者規模以外を合併、従業者規模は分割（30～49 人に限定）

「82A 普通洗濯業、全経営組織、全事業所、30～49 人、全地域」（n=262）



それぞれの合併の場合におけるパラメータの推定値を表２－４に示す。網掛けは、5%水準で有意でないことを表す。従業者規模の合併により、雇用者数の係数がかなり小さくなることが分かる。

表２－４　層の合併による推定値の変化

変数変換	定数項の有無	パラメータ	合併しない	地域を合併	経営組織の合併	従業者規模を合併	従業者規模以外を合併	従業者規模を分割
変換なし	なし	雇用者数	205.10	201.83	230.11	182.67	244.81	236.09
	あり	定数項	2217.10	-432.44	351.81	-29.90	-258.66	1979.22
		雇用者数	145.60	213.79	224.34	185.50	247.52	185.14
対数変換		定数項	5.88	4.85	5.47	-	5.17	7.65
		雇用者数(対数)	0.84	1.13	0.98	-	1.03	0.35
サンプルサイズ			5	8	46	1108	483	262

この分析では、産業小分類の合併は比較していない。同一の中分類の中で１雇用者当たり給与支給額の水準が大きく異なることがあり、中分類内での同質性・異質性を十分に確認する必要があるだろう。

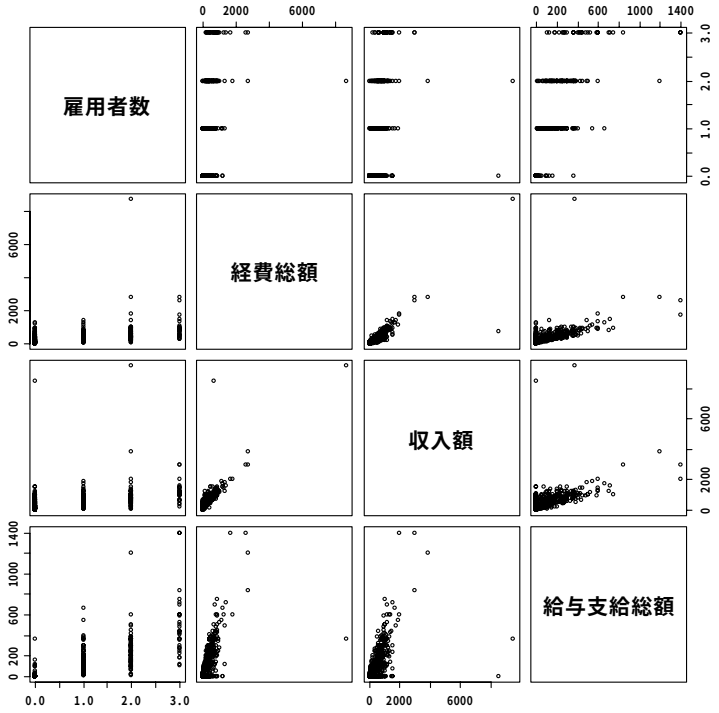
（４）多変量への拡張

ここまでは、雇用者数を説明変数とする１変量モデルを見てきた。米国の経済センサスの実験では、定数項のあるモデルのほかに、複数の説明変数があるモデルの有効性が示されている（Williams and Thompson, 2004、Thompson and Williams, 2003）。そこで、次に、複数の説明変数によるモデルを検討する。

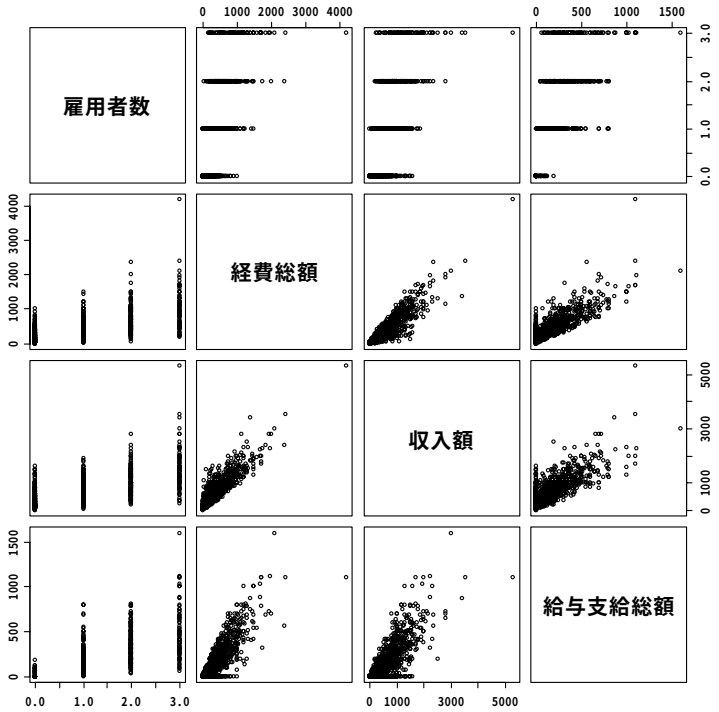
給与支給総額に対して、雇用者数、経費総額、収入額の３つを説明変数の候補として、２変数ずつ組み合わせた散布図（図２－４）を示す。図から、変数間におおむね強い関連が見られるが、線形性の高い場合や非線形性が疑われる場合など、層によって特徴が異なることが分かる。

図２－４　典型層における雇用者数、経費総額、収入額の散布図

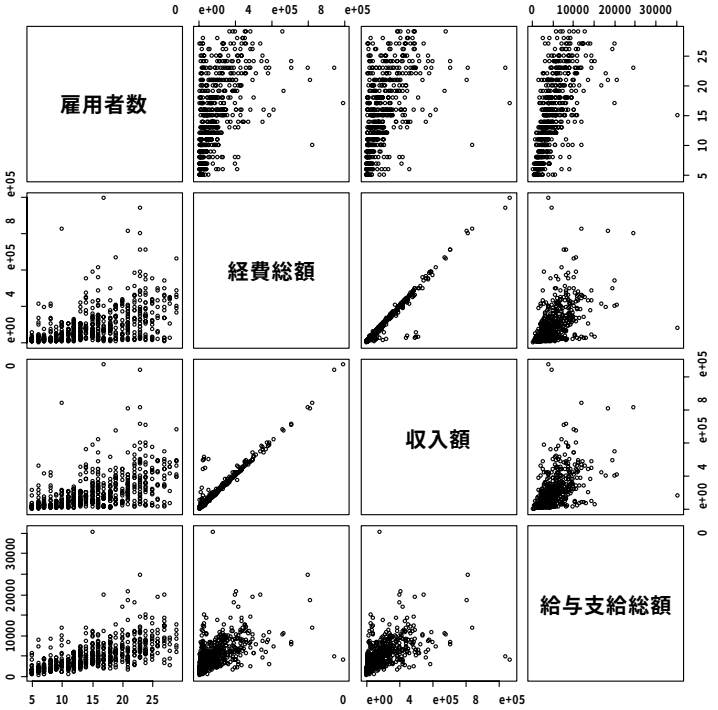
①「82A 普通洗濯業、個人経営、単独事業所、0～4 人、人口 30 万未満市」



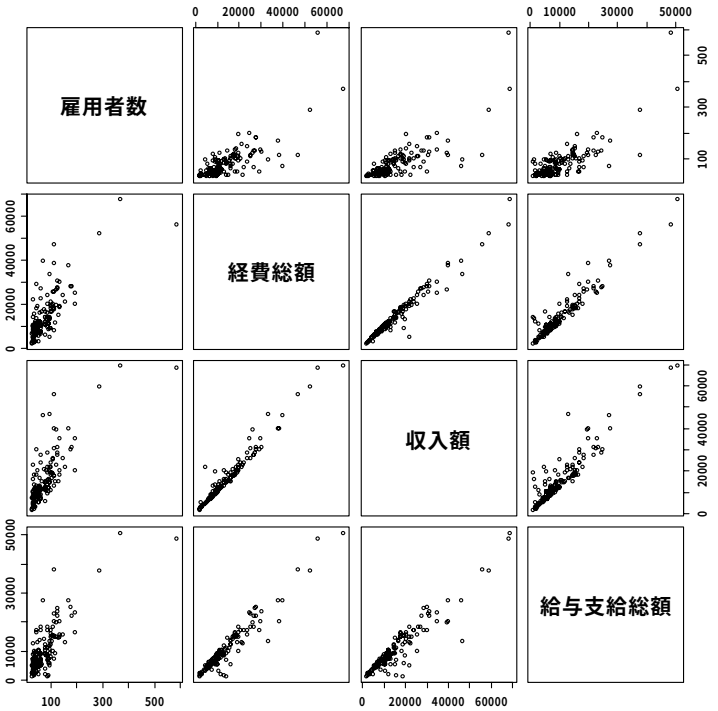
②「823 美容業、個人経営、単独事業所、0～4 人、人口 30 万未満市」



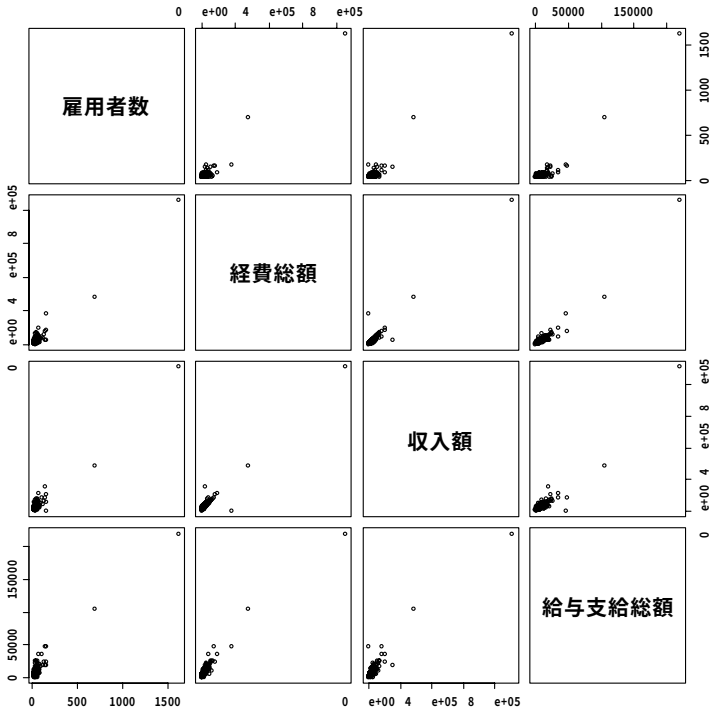
③「84J パチンコホール、会社、単独事業所、5～29 人、人口 30 万未満市」



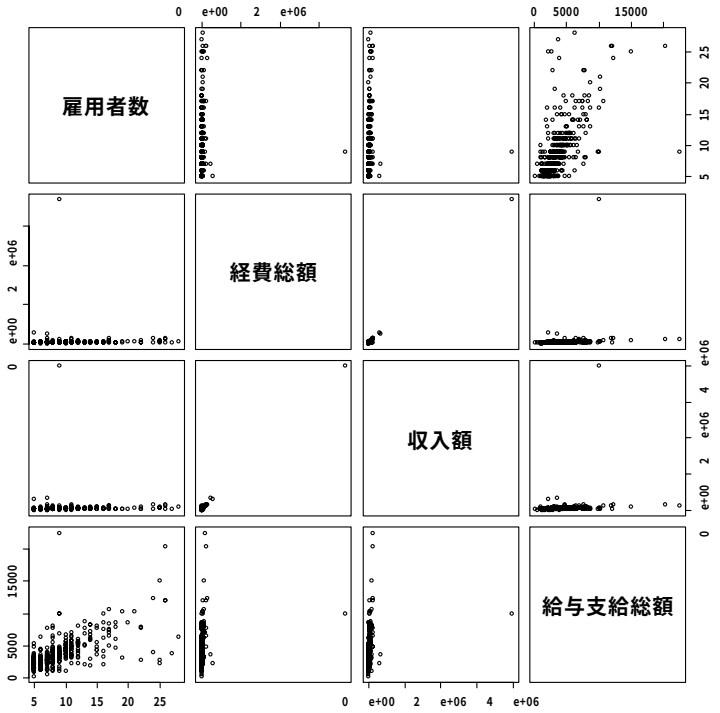
④「75H 訪問介護事業、会社、単独事業所、30 人以上、県庁所在市・人口 30 万以上市」



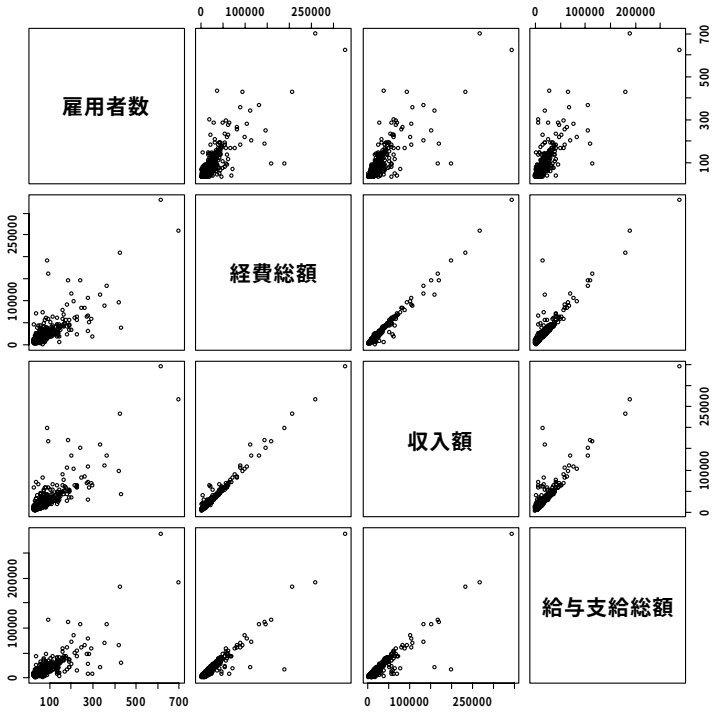
⑤「70A 一般食堂、会社、支所･支社･支店、30 人以上、県庁所在市・人口 30 万以上市」



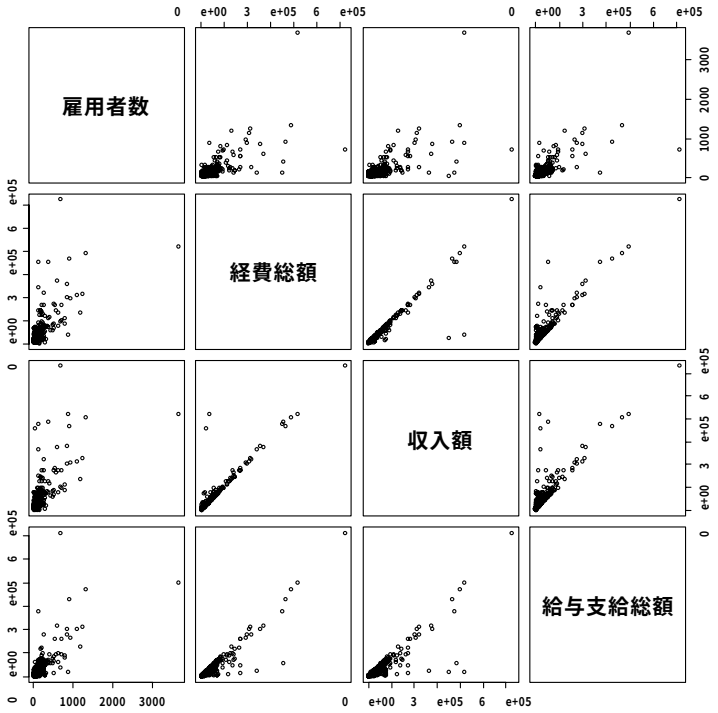
⑥「882 産業用機械器具賃貸業、会社、支所･支社･支店、5～29 人、人口 30 万未満市」



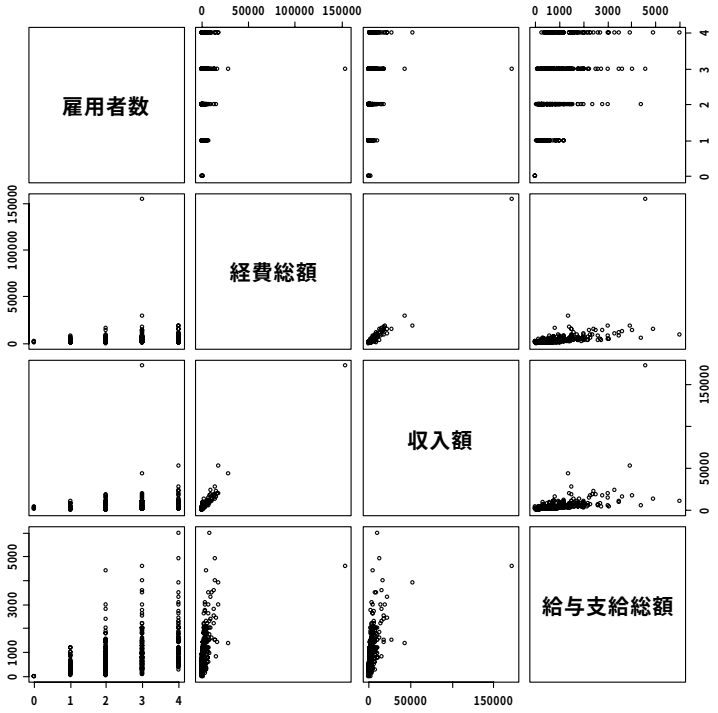
⑦「906 警備業、会社、単独事業所、30 人以上、県庁所在市・人口 30 万以上市」



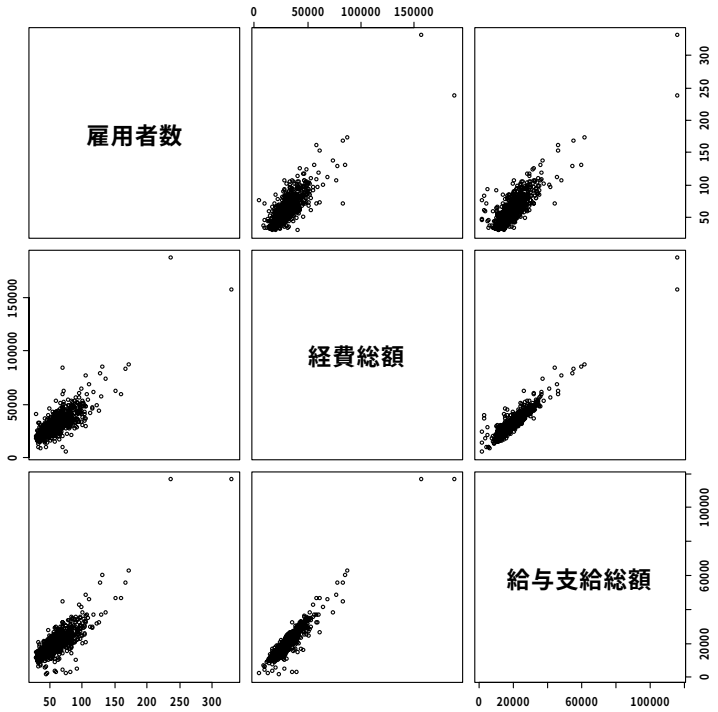
⑧「90A 労働者派遣業、会社、単独事業所、30 人以上、県庁所在市・人口 30 万以上市」



⑨「691 不動産賃貸業（貸家業，貸間業を除く）、会社、単独事業所、0～4 人、県庁所在市・人口 30 万以上市」



⑩「75C 特別養護老人ホーム、その他、単独事業所、30 人以上、町村」



（５）予備的分析からの考察

ア　線形回帰モデルの拡張

前記（２）の１変量モデルの当てはめにおいて、以下の傾向が見られた。

- 従業者規模 0～4 人の層は、現行の方法と同等の定数項がないモデルの当てはまりが良い傾向がある。一方で、説明変数の取り得る値が離散で少ないため（０、１、２、３、４）、回帰モデルの当てはめに注意が必要である。さらに、雇用者数の大きさによって給与支給総額の分散が変わる分散不均一の傾向が見られるため、平方根変換を検討する必要があると思われる。
- 従業者規模 5～29 人の層は、定数項のあるモデルの当てはまりが良い場合がある。
- 従業者規模 30 人以上の層は、定数項のあるモデルや説明変数、被説明変数ともに対数をとるモデル（乗法モデル）の適用を検討する必要があると思われる。

これらの結果からは、定数項や変数変換の導入が有効であると考えられる。また、散布図から、外れ値や分散不均一に対しては、比推定型回帰の一般化や、LTS（Least Trimmed Squares）などのロバスト回帰の適用が考えられる。

一方で、従業者規模による層化と、雇用者数を説明変数とする回帰モデルの併用は、従業者規模階級の端で回帰予測値が不連続になるという問題が起きる。

イ　多変量への拡張

多変量への拡張を試みたが、利用できる説明変数が限られていることに加えて、（４）の散布図で示したように変数間に強い相関が認められる。説明変数間の相関から生じる多重共線性を回避するためには、リッジ回帰の適用が有効であるかもしれない。

しかし、経理項目の３つの変数は同時に欠測となる傾向があるため、経理項目以外の変数を説明変数とする回帰モデルの方が、補定対象範囲は広く、利便性が高いと考えられる。

ウ　層の再構築の検討

予備的分析では、経費総額に占める給与支給総額の割合、提供対象による分類などを参考にして、恣意的に典型的な産業の選定を行った。サンプルサイズの小さい細かい層を合併する方法とは逆に、大きな層を分割する方法も考えられる。例えば、中分類内での同質性・異質性を指標にして、多段層別分析による分割を行う方法がある。

３．回帰モデルの拡張

予備的分析では、現行の補定法と同様の層別で、雇用者数を説明変数とする給与支給総額の回帰モデルを検証した。分析に選んだ典型的な層の例では、定数項のあるモデルや変数変換の必要性を示す結果が散見された。また、サンプルサイズの小さい層が多くあることから、安定した推定に必要なサンプルサイズを確保するために、層の合併が必要である。

多変量への対応を含む種々のモデルを検討対象としたが、その中で、線形回帰モデルを拡張した２種類の回帰モデルについて、さらに検討を進める。以下に、２種類のモデルの当てはめと、補定の観点からの評価の方法を説明する。

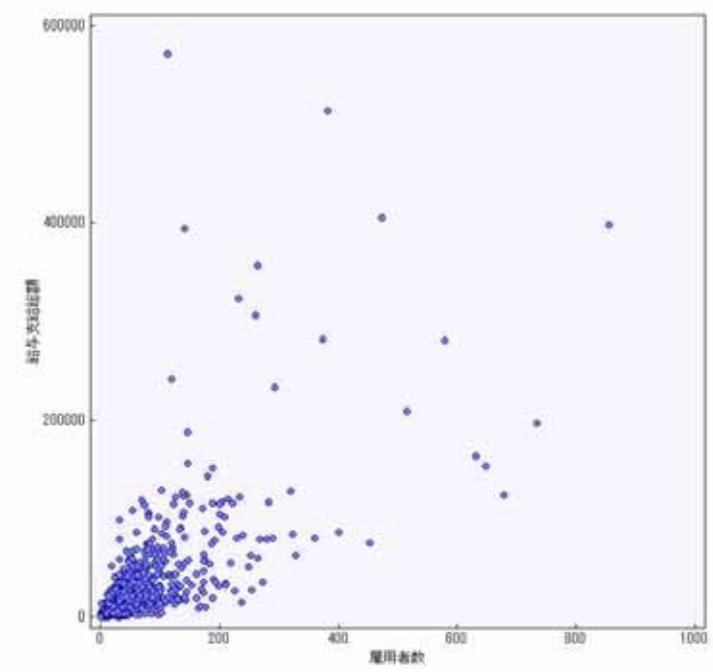
（１）２つの回帰モデル

これまで見たように、給与支給総額Yと雇用者数Xの散布図から、線形回帰モデル$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$の誤差項は、分散不均一が示唆される場合がある。図３－１に、例を再掲する。図から、雇用者数が大きくなるにしたがって、給与支給総額のばらつきも大きくなっていることが分かる。

図３－１　分散不均一の例

①中分類 69 不動産賃貸業・管理業

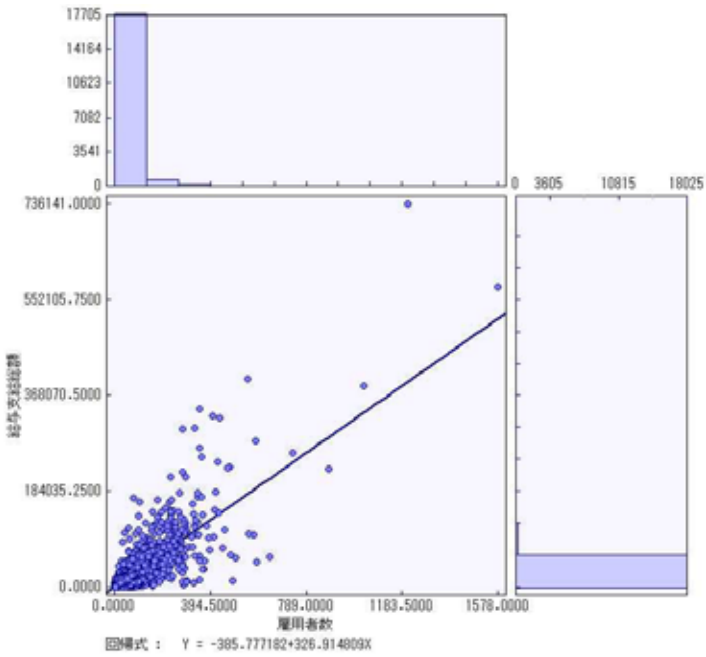
X 軸：雇用者数（単位：人），Y 軸：給与支給総額（単位：万円）



項目	横軸	縦軸
変数番号	7	6
変数名	雇用者数	給与支給総額
データ数	8743	8743
最小値	0.000	0.000
最大値	856.000	572200.000
平均値	8.8509	3753.3822
標準偏差	33.16752	18315.98078
相関係数	0.748	

②大分類 N 医療，福祉

X 軸：雇用者数（単位：人），Y 軸：給与支給総額（単位：万円）



項目	横軸	縦軸
変数番号	7	6
変数名	雇用者数	給与支給総額
データ数	18280	18280
最小値	0.000	0.000
最大値	1578.000	736141.000
平均値	42.4411	13488.8414
標準偏差	47.41761	18441.63384
相関係数	0.841	
回帰定数	-385.777	
回帰係数1次	326.915	
t 値	209.788	
P値（両側）	0.000	

ここで、分散不均一の対策として、次の２つのモデルを取り上げる。

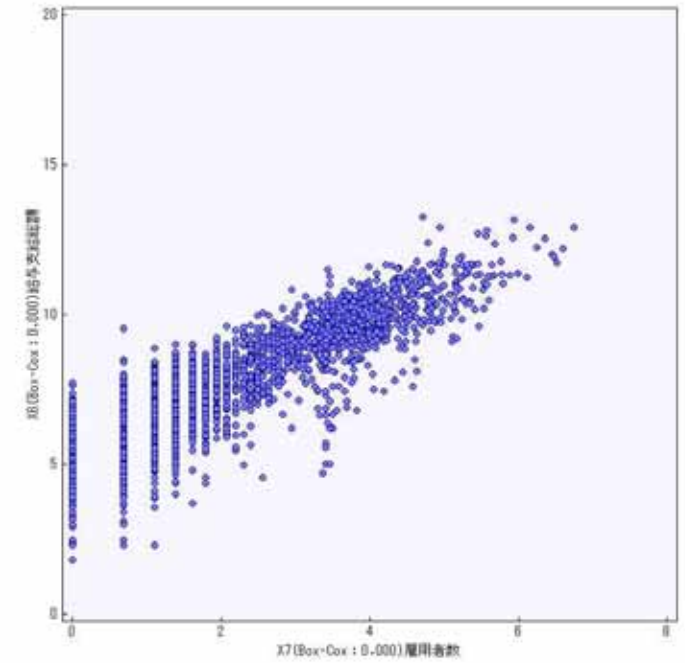
ア　対数線形回帰モデル

$$\log y_i = \beta_0 + \beta_1 \log x_i + \varepsilon_i$$

これは、$y_i = \exp(\beta_0) \cdot x_i^{\beta_1} \cdot \exp(\varepsilon_i)$と表せるので、積モデルである。また、$\varepsilon$が小さいとき、$\exp(\varepsilon)$は$1+\varepsilon$で近似できるので、誤差項が$y_i$の期待値に比例する。$\varepsilon$に等分散性を仮定すると、積モデルでは相対誤差の等分散性を仮定したといえる。

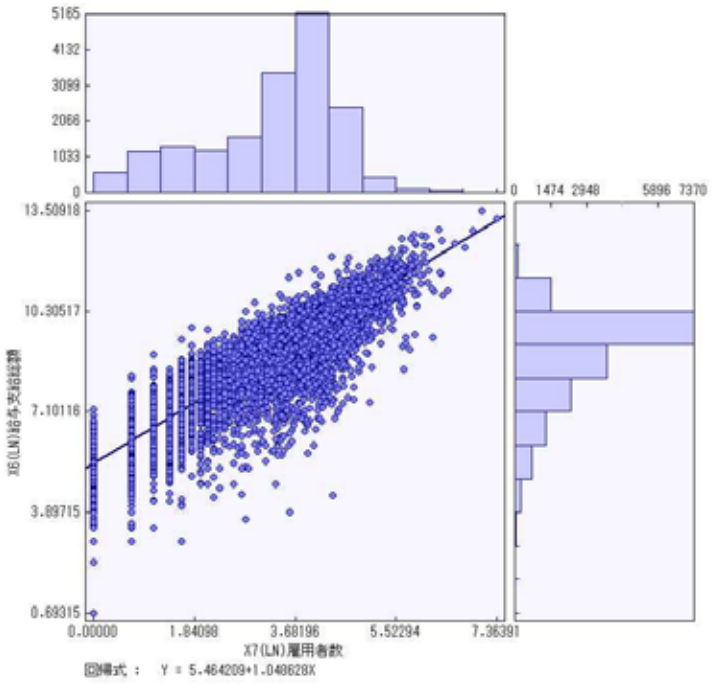
図３－２　対数変換後の散布図

①中分類 69 不動産賃貸業・管理業



項目	横軸	縦軸
変数番号	18	17
変数名	X7(Box-Cox：0.0	X8(Box-Cox：0.0
データ数	4463	4463
最小値	0.0000	1.7918
最大値	6.7523	13.2572
平均値	1.59450	7.05075
標準偏差	1.417358	1.848955
相関係数	0.887	

②大分類 N 医療，福祉



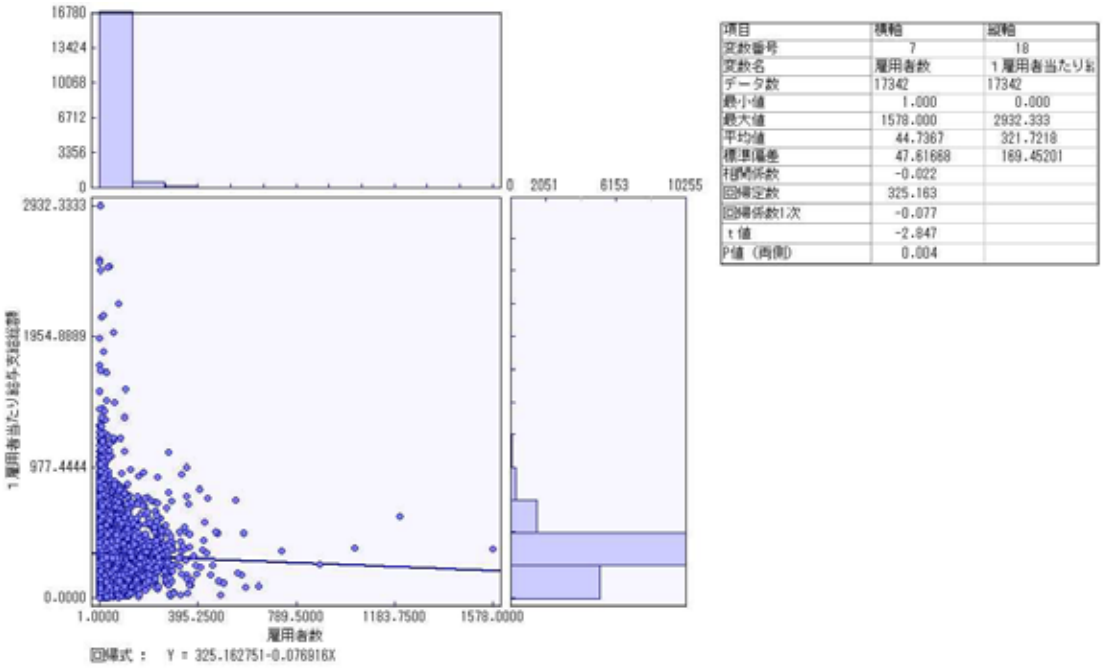
項目	横軸	縦軸
変数番号	17	16
変数名	X7(LN)雇用者数	X6(LN)給与支給総
データ数	17340	17340
最小値	0.0000	0.6931
最大値	7.3639	13.5092
平均値	3.24082	8.86282
標準偏差	1.251946	1.450107
相関係数	0.905	
回帰定数	5.464	
回帰係数1次	1.049	
t 値	280.682	
P値（両側）	0.000	

イ　１人当たり回帰モデル（原単位回帰モデル）

$$\frac{y_i}{x_i} = \beta_0 + \beta_1 x_i + \varepsilon_i$$

これは、$y_i = \beta_0 x_i + \beta_1 x_i^{\,2} + x_i \cdot \varepsilon_i$ と表され、y_i から見ると、誤差項が説明変数に比例する。また、β_1 が正のときは、雇用者数（規模）が大きくなると１人当たり支給額が増え、負のときは、雇用者数（規模）が大きくなると１人当たり支給額が減ることを意味している。

図３－３　雇用者数と１雇用者当たり給与支給総額の散布図
大分類 N 医療，福祉



予備的分析では、現行の補定法を参考に、従業者規模の階級別に推定することによって分散不均一に対処した結果を示した。しかし、階級別の推定は、各階級の端で回帰モデルの断層が生じる欠点がある。上記の２つの方法は、従業者規模による階級分けを使わずに雇用者数との関係をモデル化するため、断層は生じない。ただし、雇用者数が小さいところでは注意が必要であることを、予備的分析で指摘した。これは、説明変数である雇用者数の取り得る値が離散で少なく、給与支給総額のばらつきが相対的に大きいためである。回帰モデルの当てはめに説明変数自体が持つ誤差を考慮すると、回帰補定に代えて、平均値補定等の別途の処理が有効かもしれない。

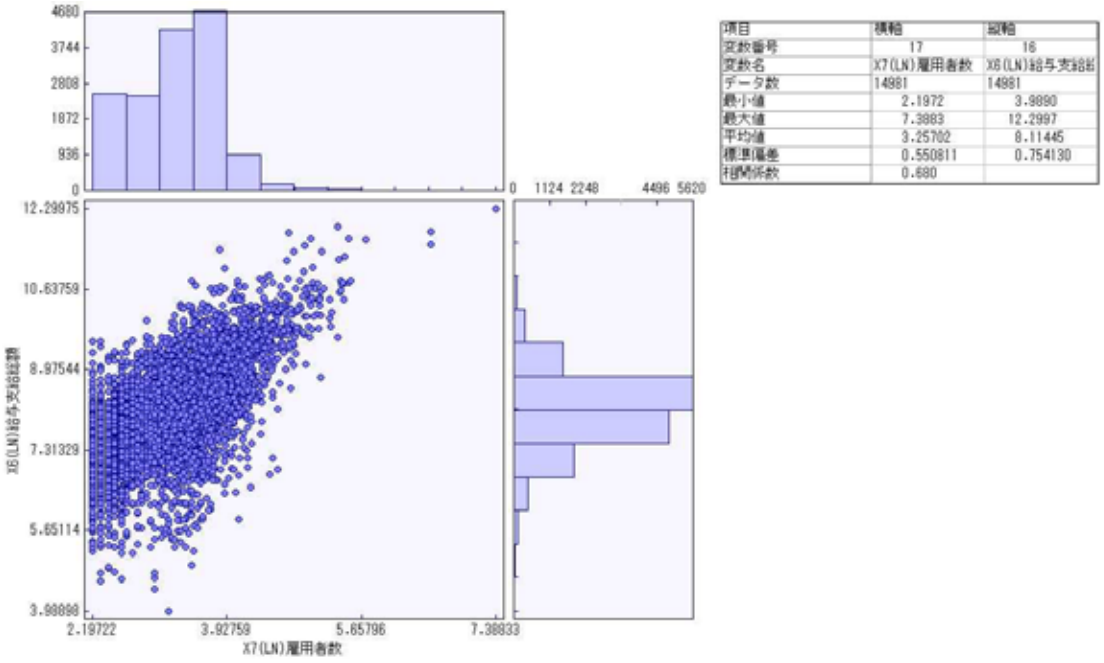
（2）モデルの選択

この節では、補定の観点から予測残差を使ってモデルを比較する方法を説明する。

まず、中分類「70 一般飲食店」を例にして、対数線形回帰モデルと１人当たり回帰モデルの当てはめを説明する。なお、雇用者数が小さいところでは、給与支給総額のばらつきが大きいことが確認されたため、回帰モデルの当てはめは、雇用者数 9 人以上に限定した。図３－４に雇用者数と給与支給総額の散布図を示す。散布図からは、雇用者数が非常に多い３レコードの存在が確認できるので、以下のモデルの当てはめからは除外する。

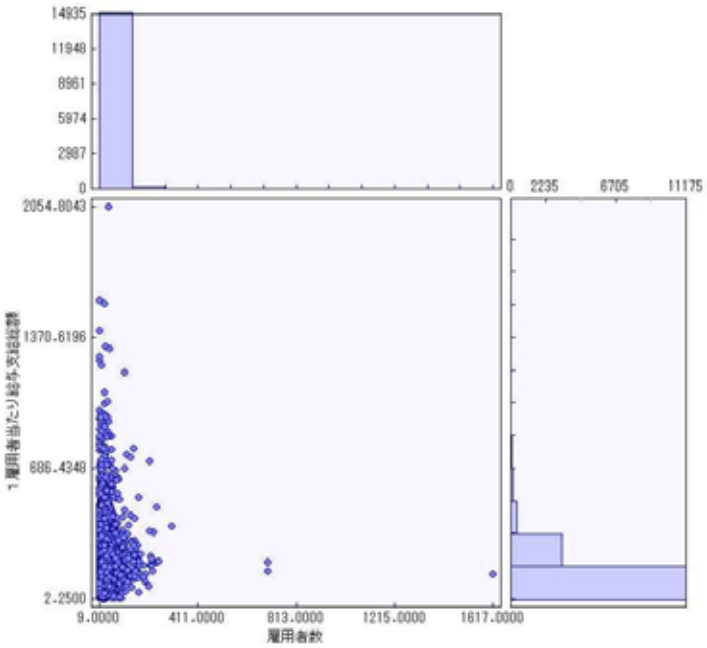
図３－４　中分類「70 一般飲食店」における雇用者数と給与支給総額

①対数変換後



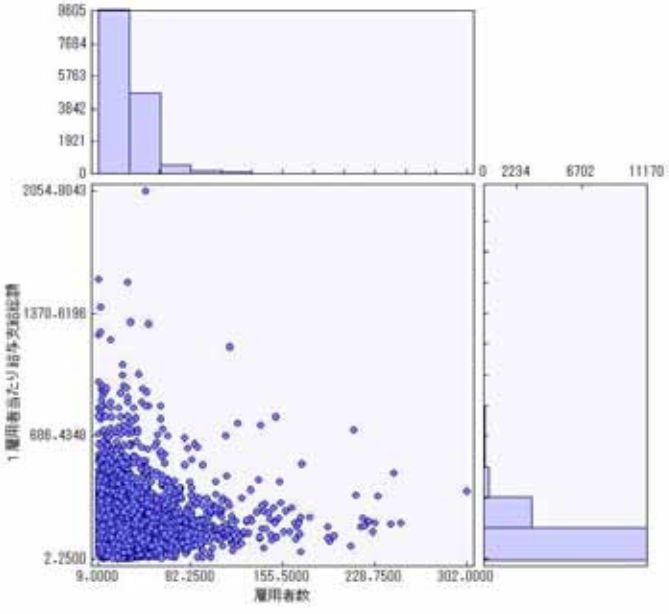
項目	横軸	縦軸
変数番号	17	16
変数名	X7(LN)雇用者数	X6(LN)給与支給総
データ数	14981	14981
最小値	2.1972	3.9890
最大値	7.3883	12.2997
平均値	3.25702	8.11445
標準偏差	0.550811	0.754130
相関係数	0.680	

②雇用者数と１雇用者当たり給与支給総額



項目	横軸	縦軸
変数番号	7	25
変数名	雇用者数	１雇用者当たり給
データ数	14881	14881
最小値	9.000	2.250
最大値	1617.000	2054.804
平均値	30.2639	149.9811
標準偏差	23.44264	99.43678
相関係数	-0.003	

③雇用者数が大きい３レコード除外後の雇用者数と１雇用者当たり給与支給総額



項目	横軸	縦軸
変数番号	7	25
変数名	雇用者数	１雇用者当たり給
データ数	14878	14878
最小値	9.000	2.250
最大値	302.000	2054.804
平均値	30.0890	149.9789
標準偏差	17.95131	99.44591
相関係数	-0.004	

前述の２つのモデルを基にして、現行の補定法で層化変数としている産業小分類、経営組織、本所・支所の別、地域を説明変数に加える。いずれも質的変数であるため、量的変数である雇用者数と質的変数が混在した回帰分析を行う。つまり、重回帰分析と数量化Ｉ類を併用したモデルとなる。

表３－１　対数線形回帰モデルの当てはめ
目的変数：対数給与支給総額
説明変数：対数雇用者数、産業小分類、経営組織、本所・支所の別、地域

目的変数名	残差平方和	重相関係数	寄与率 R^2	R*^2
対数給与支給総額	3879.645	0.736	0.542	0.542
	R**^2	残差自由度	残差標準偏差	
	0.542	14967	0.509	
説明変数名	残差平方和	変化量	分散比	偏回帰係数
定数項	10744.310	6864.662	26482.6800	4.891
産業小分類 (統合)	4175.288	295.643	228.1078	
70A,70C				0.000
70B,703				0.128
70D				0.065
70E,702,704				-0.051
70F,70H,70J				-0.146
70G				-0.403
経営組織 (統合)	3990.094	110.449	426.0924	
個人				0.000
会社, その他				0.440
本所･支所の別	4130.139	250.494	483.1815	
単独				0.000
本所				0.270
支所				-0.200
地域 (統合)	3918.397	38.752	149.4988	
人口 30 万以上市				0.000
30 万未満市, 町村				-0.103
対数雇用者数	7075.649	3196.004	12329.6300	0.923

ここで、中分類「70 一般飲食店」は、以下の小分類に分けられている。

符号	分類名
70	一般飲食店
701	食堂，レストラン
70A	一般食堂
70B	日本料理店
70C	西洋料理店
70D	中華料理店
70E	焼肉店（東洋料理のもの）
70F	その他の食堂，レストラン
702	そば・うどん店
703	すし店
704	喫茶店
709	その他の一般飲食店
70G	ハンバーガー店
70H	お好み焼店
70J	他に分類されない一般飲食店

なお、質的変数の偏回帰係数（カテゴリースコアと呼ぶ）の差の小さいカテゴリーは統合した。カテゴリーの統合は、次節で説明する。

表３－２　１人当たり回帰モデルの当てはめ
目的変数：1 雇用者当たり給与支給総額
説明変数：雇用者数、産業小分類、経営組織、本所・支所の別、地域

目的変数名	残差平方和	重相関係数	寄与率 R^2	R*^2
1雇用者当たり給与支給総額	125892045.1	0.387	0.150	0.149
	R**^2	残差自由度	残差標準偏差	
	0.149	14967	91.713	
説明変数名	残差平方和	変化量	分散比	偏回帰係数
定数項	136668728.7	10776683.610	1281.2138	134.111
産業小分類 (統合)	131189377.4	5297332.285	125.9574	
70A,70C				0.000
70B,703				19.900
70D				11.944
70E,702,704				-6.035
70F,70J,70H				-15.883
70G				-51.513
経営組織 (統合)	127908638.7	2016593.553	239.7479	
個人				0.000
会社, その他				59.078
本所･支所の別	139257280.1	13365234.970	794.4802	
単独				0.000
本所				70.848
支所				-40.409
地域 (統合)	127200492.5	1308447.423	155.5581	
人口 30 万以上市				0.000
30 万未満市, 町村				-18.857
雇用者数	125981712.7	89667.609	10.6604	-0.143

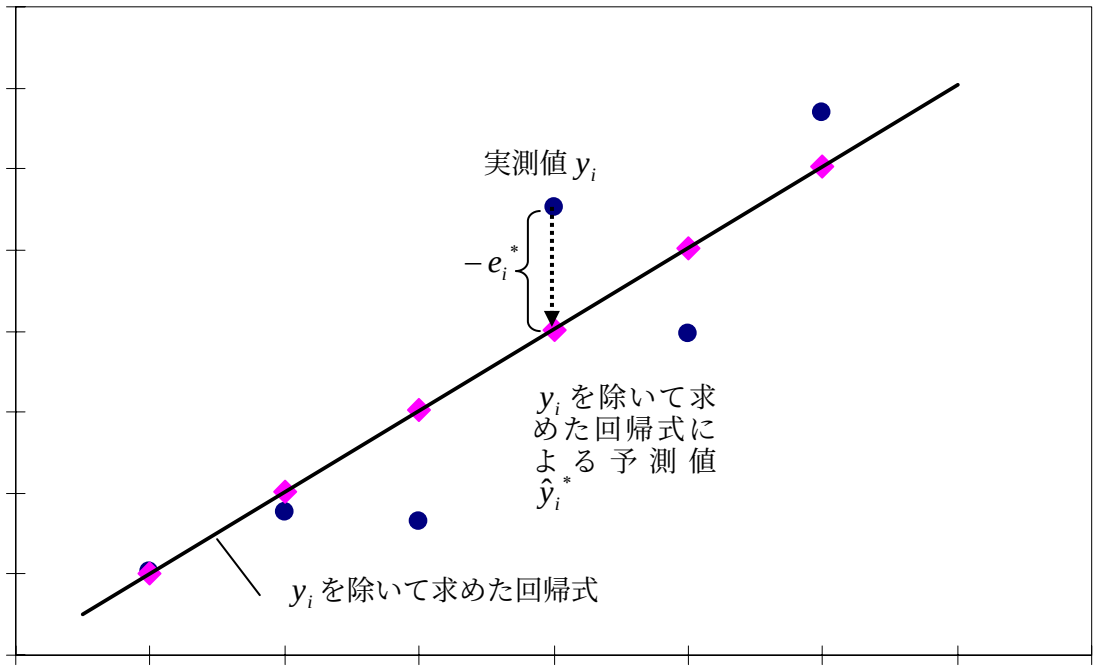
ここで、雇用者数の偏回帰係数は負であり、規模が大きいほど１人当たり支給額が減ることに注意する。

次に、対数線形回帰モデルと１人当たり回帰モデルの比較を行う。２つのモデルの間には階層性がなく、正規性の仮定の下での通常の仮説検定や自由度２重調整寄与率による単純な比較ができない。また、補定の観点を取り入れた評価が必要である。本研究では、予測残差を用いた比較の方法を検討した。

予測残差は、「i 番目の観測値を除いて求めた回帰式による予測値 $\hat{y}_i^{*}$ と実測値 y_i との差 e_i^{*}（つまり、$e_i^{*} = y_i - \hat{y}_i^{*}$）」と定義される。予測の観点では、予測値 $\hat{y}_i^{*}$ に誤差 e_i^{*} を加えたものが実測値 y_i であると考えるが、補定の観点からは、実際に観測された値 y_i が真の値であり、仮にそれが欠測であった場合の補定値として $\hat{y}_i^{*}$ を代用すると考える。このとき、相対予測残差を次式で定義する。相対予測残差が小さい方が、モデルの当てはまりが良いと考えられる。

$$\frac{-e_i^{*}}{y_i} = \frac{\hat{y}_i^{*} - y_i}{y_i}$$

図３－５　実測値と予測残差



対数線形回帰モデルと１人当たり回帰モデルのそれぞれを当てはめたときの相対予測残差を計算し、散布図と基本統計量を次ページに示す。

図３－６　２つのモデルの相対予測残差の比較　散布図

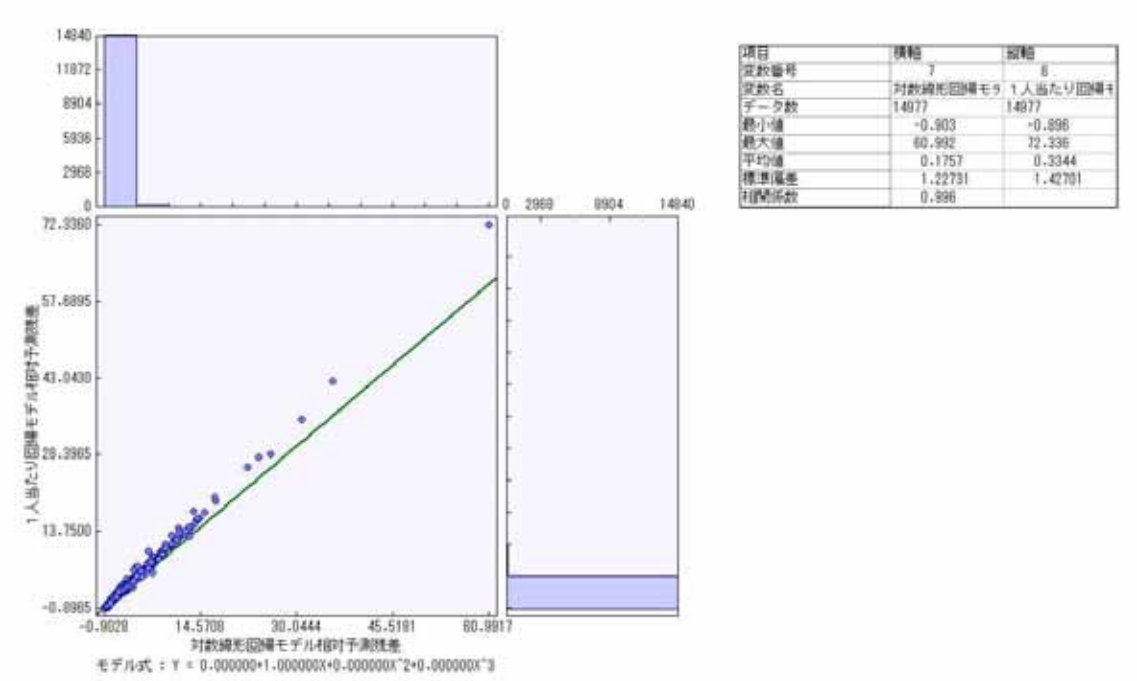


表３－３　２つのモデルの相対予測残差の比較　基本統計量

変数名	ﾃﾞｰﾀ数	合計	最小値	最大値	平均値	標準偏差	変動係数	ひずみ	とがり
対数線形回帰モデル相対予測残差	14978	2631.448	-0.903	60.992	0.1757	1.22727	6.9855	17.273	579.639
1人当たり回帰モデル相対予測残差	14978	5008.665	-0.896	72.336	0.3344	1.42697	4.2672	17.628	609.106

ひげ端(下)	1/4 分位	メジアン	3/4 分位	ひげ端(上)	外れ値数	外れ値数(下)	外れ値数(上)	メジアン信頼下限	メジアン信頼上限
-0.903	-0.237	-0.003	0.279	1.052	874	0	874	-0.011	0.004
-0.896	-0.144	0.122	0.451	1.338	873	0	873	0.113	0.131

ここで、相対予測残差の値が非常に大きくなるケースが確認されたため、ひげ端より外のレコードを除外することとする。なお、相対予測残差が非常に大きい値をとるケースは、１雇用者当たり給与支給総額の値が非常に低い場合に相当する。現行の補定法では、１雇用者当たり給与支給総額の下限値と上限値を決めて、下限値より低いケースと上限値より高いケースはレンジエラーとして補定値の算出から除外している。本研究では、外れ値の検出を併せて検討するために、レンジエラーのケースも含めて回帰モデルの当てはめを行っている。

該当するレコードを除外して、対数線形回帰モデルと１人当たり回帰モデルを当てはめた結果は、それぞれ、次のとおりである。

表３－４　外れ値除外後の対数線形回帰モデルの当てはめ

目的変数名	残差平方和	重相関係数	寄与率 R^2	R*^2	
対数給与支給総額	2217.225	0.807	0.651	0.651	
	R**^2	残差自由度	残差標準偏差		
	0.650	14045	0.397		
説明変数名	残差平方和	変化量	分散比	偏回帰係数	Exp(偏回帰係数)
定数項	8938.185	6720.960	42573.8798	5.073	159.653
産業小分類	2438.842	221.617	280.7663		
70A,70C				0.000	1.000
70B,703				0.127	1.135
70D				0.059	1.061
70E,702,704				-0.040	0.961
70F,70J,70H				-0.106	0.899
70G				-0.359	0.698
経営組織(統合)	2297.469	80.243	508.3003		
個人				0.000	1.000
会社, その他				0.397	1.487
本所･支所の別	2560.674	343.449	1087.7883		
単独				0.000	1.000
本所				0.335	1.398
支所				-0.250	0.779
地域(統合)	2255.677	38.452	243.5718		
人口 30 万以上市				0.000	1.000
30 万未満市, 町村				-0.105	0.900
対数雇用者数	5130.075	2912.850	18451.4297	0.913	

得られた回帰式から、補定値 $\hat{y}_i$ の算出式は次のとおりとなる。

$$\hat{y}_i = 159.653 \times \begin{pmatrix} 1.000 \\ 1.135 \\ 1.061 \\ 0.961 \\ 0.899 \\ 0.698 \end{pmatrix} \begin{matrix} \text{70A,70C} \\ \text{70B,703} \\ \text{70D} \\ \text{70E,702,704} \\ \text{70F,70J,70H} \\ \text{70G} \end{matrix} \times \begin{pmatrix} 1.000 \\ 1.487 \end{pmatrix} \begin{matrix} \text{個人経営} \\ \text{会社、その他} \end{matrix}$$

$$\times \begin{pmatrix} 1.000 \\ 1.398 \\ 0.779 \end{pmatrix} \begin{matrix} \text{単独} \\ \text{本所} \\ \text{支所} \end{matrix} \times \begin{pmatrix} 1.000 \\ 0.900 \end{pmatrix} \begin{matrix} \text{30万以上市} \\ \text{30万未満市、町村} \end{matrix} \times x_i^{0.913}$$

表３－５　外れ値除外後の１人当たり回帰モデルの当てはめ

目的変数名	残差平方和	重相関係数	寄与率 R^2	R*^2
1雇用者当たり給与支給総額	111715604.005	0.430	0.185	0.185
	R**^2	残差自由度	残差標準偏差	
	0.184	14045	89.186	
説明変数名	残差平方和	変化量	分散比	偏回帰係数
定数項	123142276.890	11426672.884	1436.5730	146.356
産業小分類	116214968.064	4499364.059	113.1329	
70A,70C				0.000
70B,703				20.933
70D				11.534
70E,702,704				-4.633
70F,70J,70H				-12.414
70G				-48.285
経営組織 (統合)	113542831.945	1827227.940	229.7210	
個人				0.000
会社, その他				59.532
本所･支所の別	128554273.312	16838669.306	1058.4874	
単独				0.000
本所				86.306
支所				-47.353
地域 (統合)	113091940.974	1376336.969	173.0345	
人口 30 万以上市				0.000
30 万未満市, 町村				-19.967
雇用者数	111858790.718	143186.712	18.0016	-0.187

$$\hat{y}_i = \left\{ 146.356 + \begin{pmatrix} 0.000 \\ 20.933 \\ 11.534 \\ -4.633 \\ -12.414 \\ -48.285 \end{pmatrix} \begin{matrix} \text{70A,70C} \\ \text{70B,703} \\ \text{70D} \\ \text{70E,702,704} \\ \text{70F,70J,70H} \\ \text{70G} \end{matrix} + \begin{pmatrix} 0.000 \\ 59.532 \end{pmatrix} \begin{matrix} \text{個人経営} \\ \text{会社、その他} \end{matrix} \right.$$

$$\left. + \begin{pmatrix} 0.000 \\ 86.306 \\ -47.353 \end{pmatrix} \begin{matrix} \text{単独} \\ \text{本所} \\ \text{支所} \end{matrix} + \begin{pmatrix} 0.000 \\ -19.967 \end{pmatrix} \begin{matrix} \text{30万以上市} \\ \text{30万未満市、町村} \end{matrix} - 0.187 \times x_i \right\} \times x_i$$

対数線形回帰モデルの当てはめについて、表３－１と３－４を比較すると、ひげ端より外のレコードを除外することによって、自由度２重調整寄与率（**R**^2**）が向上していることが分かる。同様に、１人当たり回帰モデルの当てはめについても、表３－２と３－５を

比較すると、自由度２重調整寄与率が向上している。また、定数項の値が大きくなり、１雇用者当たり給与支給総額が上にシフトしていることが確認できる。

相対予測残差を再計算し、散布図と基本統計量を図３－７、表３－６に示す。

図３－７　外れ値除外後の２つのモデルの相対予測残差の比較　散布図

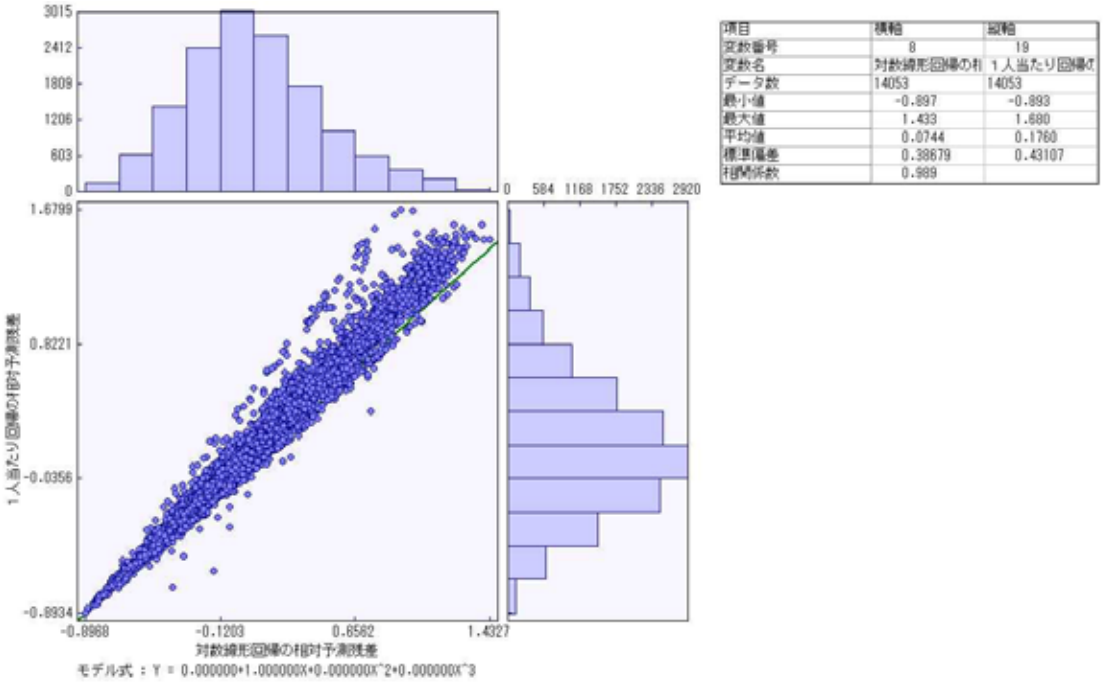


表３－６　外れ値除外後の２つのモデルの相対予測残差の比較　基本統計量

変数名	ﾃﾞｰﾀ数	最小値	最大値	平均値	標準偏差
対数線形回帰モデル相対予測残差	14053	-0.897	1.433	0.0744	0.38679
1人当たり回帰モデル相対予測残差	14053	-0.893	1.690	0.1760	0.43107

散布図は対角線の上側に点が多く、対数線形回帰モデルの相対予測残差の方が、１人当たり回帰モデルより概ね小さいことが分かる。また、レンジ、平均値、標準偏差も対数線形回帰モデルの方が小さい。

相対予測残差の小さい方がモデルの当てはまりが良いが、全体的な当てはまりの良さに基づいてモデルを比較する統計量として、次の相対予測残差の平方和を考える。

$$S_{e^*} = \sum_{i=1}^{n}\left(\frac{-e_i^{\,*}}{y_i}\right)^2 = \sum_{i=1}^{n}\left(\frac{\hat{y}_i^{\,*} - y_i}{y_i}\right)^2$$

ここで、S_{e^*}を相対予測平方和と呼ぶ。相対予測残差の平方和の小さい方が、全体的なモデルの当てはまりが良いと考えられる。

表３－７に、比較対象の対数線形回帰モデルと１人当たり回帰モデルのそれぞれについ

て、相対予測残差の平方和を示す。対数線形回帰モデルの方が小さいことが分かる。

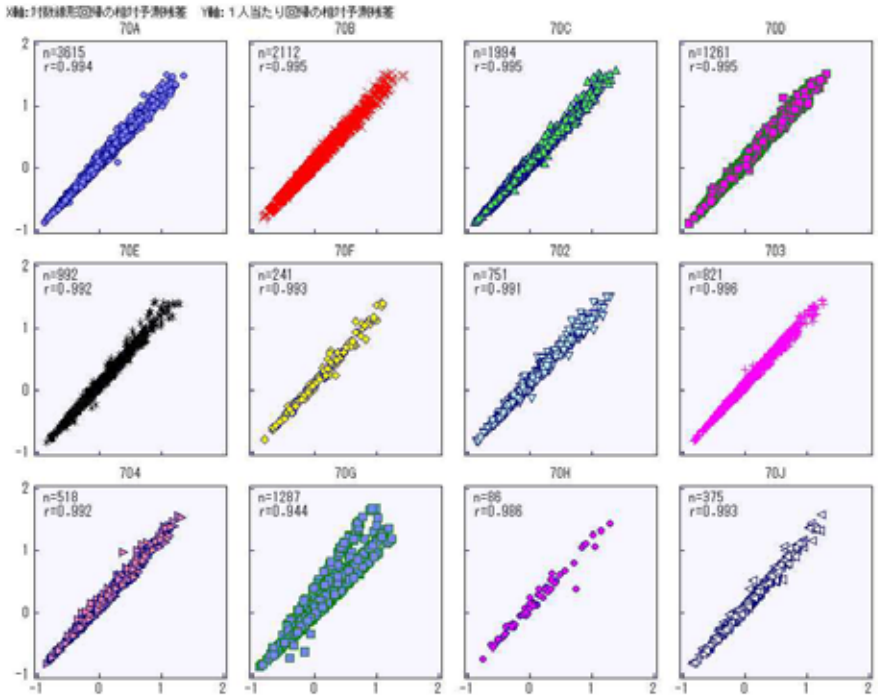
表３－７　２つのモデルの相対予測残差の平方和

種類	相対予測残差の平方和
対数線形回帰モデル	2180.005
１人当たり回帰モデル	3046.488

上記のことから、個々の相対予測残差の分布や、相対予測残差の平方和の大きさを総合的に判断して、対数線形回帰モデルの方が当てはまりが良いと考えられる。

参考に、２つのモデルの相対予測残差を産業小分類別に見ると、「70G ハンバーガー店」において２つのモデルの違いが顕著に現れることが分かる。「70G ハンバーガー店」の雇用や給与体系の情報を収集し、実態を反映したモデル化と選択方法を再検討する余地があると考えられる。また、「70 一般飲食店」から「70G ハンバーガー店」を除外して、重回帰・数量化Ｉ類モデルを適用することを検討する必要があるだろう。

図３－８　小分類別２つのモデルの相対予測残差の比較



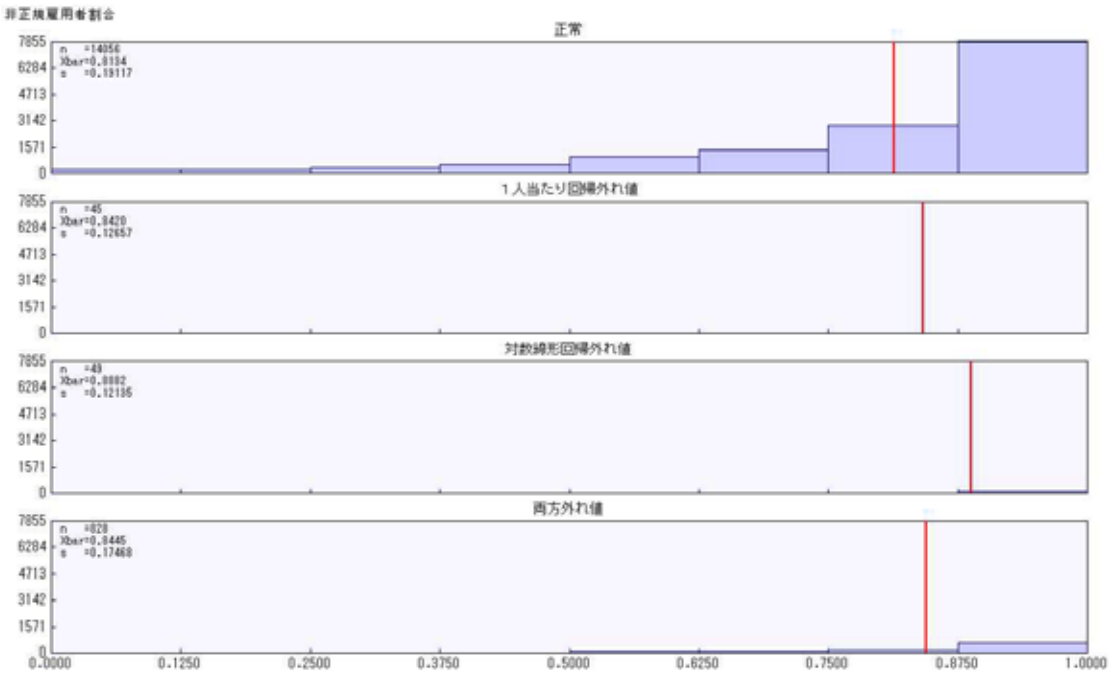
また、相対予測残差の値が非常に大きくなるケースは、分析から除外した。これらの除外したケースについて、簡単に説明する。

除外したケースは、１雇用者当たり給与支給総額が非常に低い。このことから、非正規

雇用者（常用雇用者のうちパート・アルバイトなどと、臨時雇用者を合わせたもので定義する）の割合が高いことが考えられる。図３－９に、非正規雇用者割合のヒストグラムを示す。なお、グラフ中の縦線は、非正規雇用者割合の平均を示す。除外されないケース（図中では「正常」と表示)、１人当たり回帰で除外したケース、対数線形回帰で除外したケースの平均は、それぞれ、0.81、0.84、0.89 である。つまり、除外したケースの非正規雇用者割合は、除外されないケースと比較して、若干高い傾向が見られる。

しかし、「70 一般飲食店」は、他の中分類と比較して、全体的に非正規雇用者割合が非常に高く、除外したケースが特に異常というわけではない。さらに、本研究で利用可能な範囲のデータ項目を用いて個々のケースを確認したが、除外したケースの特徴は明らかではなかった。これらの除外したケースについては、「70 一般飲食店」の背景領域の知識に基づいて、異常値であるかどうかを個別に判断する必要があるだろう。

図３－９　非正規雇用者の割合



４．層の構築

安定した推定を行うために、経理項目がすべて記入されている事業所のデータ（完全データ）のサンプルサイズが十分に大きい産業中分類を基準として、そこから細分化していく試みを説明する。ここでは、層化されていない単回帰モデルの残差を多段層別分析によって分類する方法と、現行の補定法における層化変数を説明変数に入れた重回帰・数量化Ｉ類モデルよって層化因子の効果を確認する方法を、それぞれ説明する。なお、それぞれの方法の特徴を示すため、多段層別分析による分類は中分類「69 不動産賃貸業・管理業」の事例を取り上げ、重回帰・数量化Ｉ類モデルによる層化の確認は中分類「70 一般飲食店」において１人当たり回帰モデルを当てはめた事例を取り上げる。

（１）多段層別分析による分類

まず、単回帰モデルの残差を多段層別分析によって分類する方法を示す。例として、中分類「69 不動産賃貸業・管理業」を使う。雇用者数のみを説明変数とする対数線形回帰モデルの当てはめは、次のとおりである。なお、回帰モデルの当てはめは、雇用者数 9 人以上に限定した。

表４－１　対数線形回帰モデルの当てはめ

目的変数名	残差平方和	重相関係数	寄与率 R^2	R*^2
対数給与支給総額	921.412	0.717	0.514	0.513
	R**^2	残差自由度	残差標準偏差	
	0.513	1215	0.871	

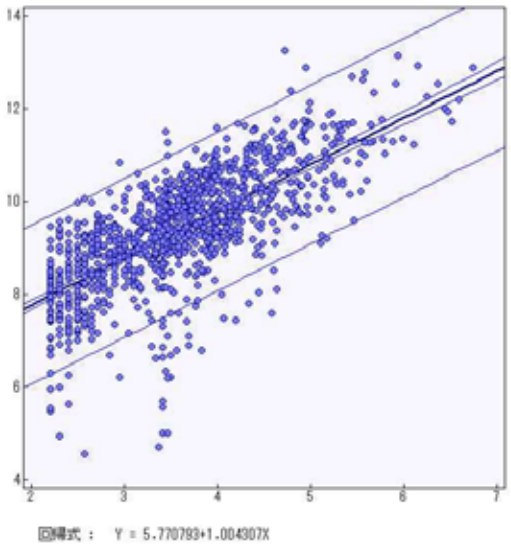
説明変数名	残差平方和	変化量	分散比	偏回帰係数
定数項	3278.868	2357.456	3108.6075	5.771
対数雇用者数	1894.08	972.668	1282.587	1.004

求められた回帰式は、

対数給与支給総額＝5.771＋1.004×対数雇用者数

となる。図４－１に、散布図に回帰式を重ねて示す。

図４－１　対数線形回帰モデルと実測値



項目	横軸	縦軸
変数番号	17	16
変数名	X7(LN)雇用者数	X6(LN)給与支給総
データ数	1217	1217
最小値	2.1972	4.5539
最大値	6.7523	13.2572
平均値	3.58193	9.36814
標準偏差	0.890531	1.248051
相関係数	0.717	
回帰定数	5.771	
回帰係数1次	1.004	
t 値	35.813	
P値（両側）	0.000	

ここで、予測値 $\hat{y}_i$ と実測値 y_i との差 e_i（つまり、$e_i = y_i - \hat{y}_i$）に着目する。残差は雇用者数だけでは説明できない部分に相当するので、他の因子の影響を検討する。他の因子として、現行の補定法の層化変数である産業小分類、経営組織、本所・支所の別、地域を使い、多段層別分析による分類を試みた。

多段層別分析は、AID（Automatic Interaction Detector）の一種であり、目的変数を最も良く説明できる変数を逐次的に２分割して、多段階で層別するものである。分割は、層間と層内の分散比が最大となるカテゴリーの組み合わせを探すことによって行われる。

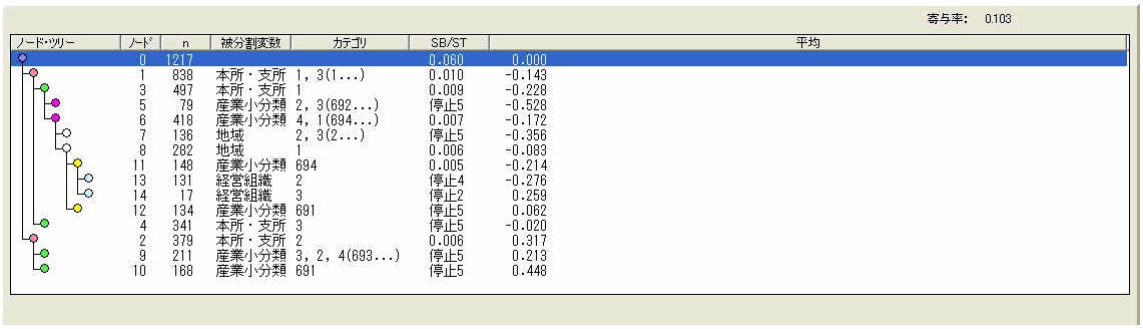
各説明変数のカテゴリーは表４－２のとおりである。いずれの変数も、順序のない質的変数である。

表４－２　カテゴリーの一覧

変数名	順序	カテゴリーと符号			
産業小分類	無	不動産賃貸業 691	貸家業，貸間業 692	駐車場業 693	不動産管理業 694
経営組織	無	個人経営 1	会社 2	その他 3	
本所･支所の別	無	単独 1	本所 2	支所 3	
地域	無	30 万以上市 1	30 万未満市 2	町村 3	

逐次的な 2 分割の結果を図４－２に示す。

図４－２　逐次的な 2 分割


寄与率:　0.103

ノード・ツリー	ノード	n	被分割変数	カテゴリ	SB/ST	平均
	0	1217			0.060	0.000
	1	838	本所・支所	1, 3(1...)	0.010	-0.143
	3	497	本所・支所	1	0.009	-0.228
	5	79	産業小分類	2, 3(692...)	停止5	-0.528
	6	418	産業小分類	4, 1(694...)	0.007	-0.172
	7	136	地域	2, 3(2...)	停止5	-0.356
	8	282	地域	1	0.006	-0.083
	11	148	産業小分類	694	0.005	-0.214
	13	131	経営組織	2	停止4	-0.276
	14	17	経営組織	3	停止2	0.259
	12	134	産業小分類	691	停止5	0.062
	4	341	本所・支所	3	停止5	-0.020
	2	379	本所・支所	2	0.006	0.317
	9	211	産業小分類	3, 2, 4(693...)	停止5	0.213
	10	168	産業小分類	691	停止5	0.448

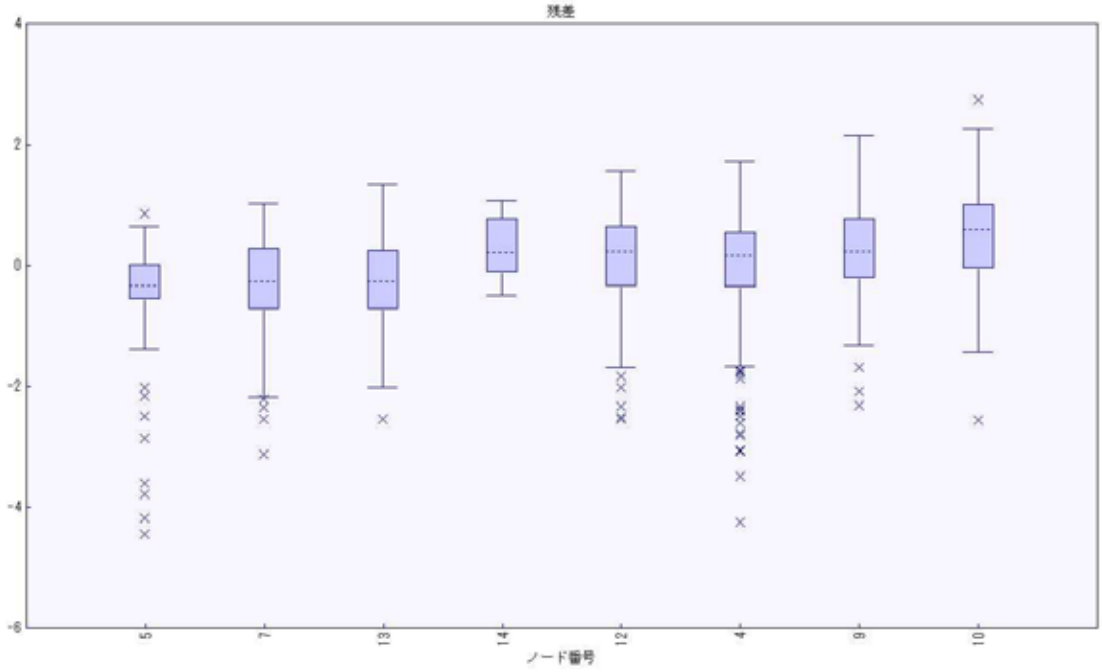
最初に、本所・支所の別によって、「単独（1)」と「支所（3)」の組み合わせのノード 1 と、「本所（2)」のノード 2 に 2 分割される。ノード 1 は、同じく本所・支所の別によって、「単独」がノード 3 に、「支所」がノード 4 に 2 分割される。ノード 3 は、次に、産業小分類によって、「692 貸家業，貸間業」と「693 駐車場業」の組み合わせのノード 5 と、「691 不動産賃貸業」と「694 不動産管理業」の組み合わせのノード 6 に 2 分割される。ノード 4 は、それ以上は分割されない。最終的に、８つのノードに分割され、それぞれの分類の組み合わせは、表４－３のとおりである。

表４－３　ノードによる分類

ノード	産業小分類	経営組織	本所・支所の別	地域
ノード 5	692、693	すべて	単独	すべて
ノード 7	691、694	すべて	単独	30 万未満市、町村
ノード 13	694	会社	単独	30 万以上市
ノード 14	694	その他	単独	30 万以上市
ノード 12	691	すべて	単独	30 万以上市
ノード 4	すべて	すべて	支所	すべて
ノード 9	692、693、694	すべて	本所	すべて
ノード 10	691	すべて	本所	すべて

図４－３の箱ひげ図でノード別残差の分布を確認すると、平均の位置の違いに特徴が見られる。

図４－３　ノード別残差の箱ひげ図



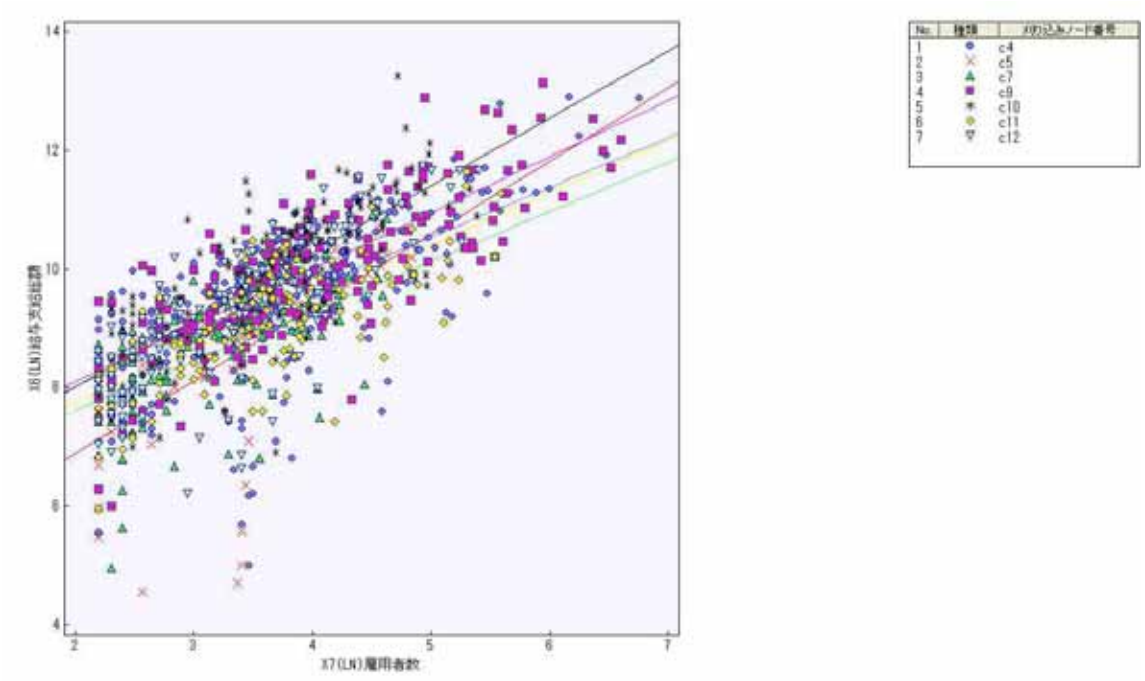
ここで、ノード 14 は n=17 と小さいため（図４－２）、ノード 11 まで刈り込むことにした。つまり、ノード 13 と 14 を合併することになる。刈り込んだ結果のノード別残差の基本統計量を表４－４に示す。

表４－４　刈り込み後ノード別残差の基本統計量

	ノード 4	ノード 5	ノード 7	ノード 9	ノード 10	ノード 11	ノード 12
データ数	341	79	136	211	168	148	134
最小値	-4.248	-4.452	-3.127	-2.320	-2.570	-2.552	-2.543
最大値	1.712	0.859	1.024	2.144	2.748	1.335	1.575
平均値	-0.020	-0.528	-0.356	0.214	0.448	-0.214	0.062
標準偏差	0.914	1.027	0.786	0.727	0.792	0.709	0.832
ひずみ	-1.393	-2.269	-0.966	-0.265	-0.470	-0.452	-1.012
とがり	2.538	5.473	1.068	0.377	0.773	0.217	1.022

次に、ノード別に対数線形回帰モデルを当てはめ直し、分類の効果を確認する。ノード別の散布図を図４－４に示す。また、基本統計量、回帰係数を表４－５に示す。

図４－４　ノード別対数線形回帰モデルの当てはめ



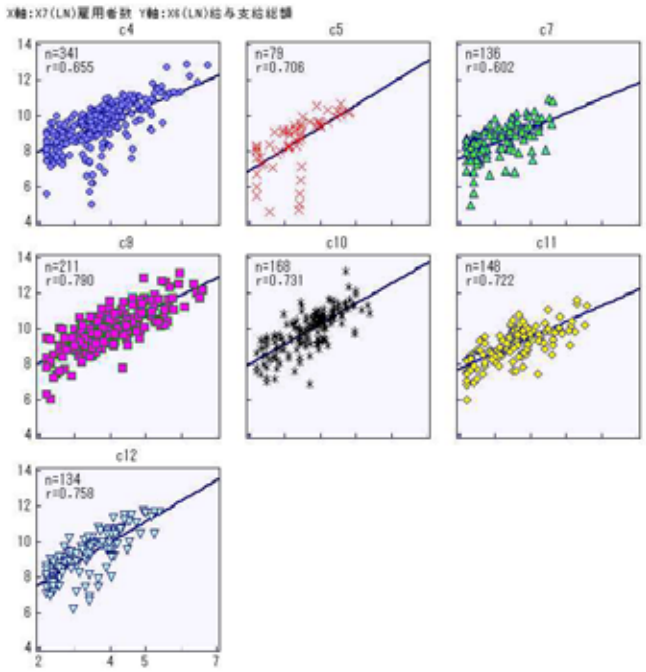


表４－５　ノード別基本統計量、回帰係数

	ノード 4	ノード 5	ノード 7	ノード 9	ノード 10	ノード 11	ノード 12
データ数	341	79	136	211	168	148	134
Y最大値	12.913	10.671	10.926	13.149	13.257	11.651	11.747
Y平均値	9.357	8.838	8.563	9.956	9.912	9.136	9.176
Y最小値	5.004	4.554	4.956	6.004	6.906	5.956	6.215
Y標準偏差	1.193	1.427	0.974	1.183	1.153	1.013	1.259
相関係数	0.655	0.706	0.602	0.790	0.731	0.722	0.758
回帰定数項	6.341	4.418	5.931	6.204	5.737	5.989	5.286
回帰係数	0.840	1.235	0.840	0.949	1.135	0.883	1.168
t値	15.976	8.755	8.721	18.641	13.819	12.594	13.355
P 値 (両側)	0.000	0.000	0.000	0.000	0.000	0.000	0.000

ノード 5（「692 貸家業，貸間業・693 駐車場業、単独事業所」）、10（「691 不動産賃貸業、本所事業所」）、12（「691 不動産賃貸業、単独事業所、人口 30 万以上市」）は、他のノードよりも回帰直線の傾きが急であり、回帰係数の値は 1 を超えている。これは、雇用者数（規模）が大きくなると１人当たり支給額が増えることを示唆している。他のノードは、逆に、規模が大きくなると、１人当たり支給額が減る傾向を示している。なお、図４－４で、ノード 4（「支所事業所」）、5（「692 貸家業，貸間業・693 駐車場業、単独事業所」）において顕著な縦方向に伸びる点については、「69 不動産賃貸業・管理業」の背景領域の知見に基づいた個別の検討が必要である。

（２）重回帰・数量化Ｉ類モデルによる層化因子の効果

次に、現行の補定法における層化変数を説明変数に入れた重回帰・数量化Ｉ類モデルの適用を説明する。例として、中分類「70 一般飲食店」の１人当たり回帰モデルを用いる。当てはめの結果は、以下のとおりである。

表４－６　中分類「70 一般飲食店」における１人当たり回帰モデルの当てはめ

目的変数名	残差平方和	重相関係数	寄与率 R^2	R*^2
1雇用者当たり給与支給総額	125711866.7	0.389	0.151	0.150
	R**^2	残差自由度	残差標準偏差	
	0.149	14959	91.672	
説明変数名	残差平方和	変化量	分散比	偏回帰係数
定数項	135305267.8	9593401.148	1141.5604	132.403
産業小分類	131036064.5	5324197.800	57.5954	
70A				0.000
70B				20.523
70C				3.764
70D				12.831
70E				-3.979
70F				-11.183
702				-3.937
703				22.320
704				-6.920
70G				-50.502
70H				-20.360
70J				-14.763
経営組織	127854638.0	2142771.334	127.4888	
個人経営				0.000
会社				60.239
その他				36.335
本所･支所の別	138967039.0	13255172.270	788.6452	
単独				0.000
本所				70.129
支所				-40.937
地域	127006681.4	1294814.746	77.0378	
人口 30 万以上市				0.000
人口 30 万未満市				-19.820
町村				-14.775
雇用者数	125799084.0	87217.303	10.3784	-0.142

ここで、質的変数のカテゴリーごとの偏回帰係数については、標準誤差が相対的に大きく、差が統計的に有意ではない組み合わせがあった。それらを統合した結果を次に示す。

表４－７　カテゴリー統合済みの１人当たり回帰モデルの当てはめ

目的変数名	残差平方和	重相関係数	寄与率 R^2	R*^2
1雇用者当たり給与支給総額	125892045.1	0.387	0.150	0.149
	R**^2	残差自由度	残差標準偏差	
	0.149	14967	91.713	
説明変数名	残差平方和	変化量	分散比	偏回帰係数
定数項	136668728.7	10776683.610	1281.2138	134.111
産業小分類(統合)	131189377.4	5297332.285	125.9574	
70A,70C				0.000
70B,703				19.900
70D				11.944
70E,702,704				-6.035
70F,70J,70H				-15.883
70G				-51.513
経営組織(統合)	127908638.7	2016593.553	239.7479	
個人				0.000
会社, その他				59.078
本所･支所の別	139257280.1	13365234.970	794.4802	
単独				0.000
本所				70.848
支所				-40.409
地域(統合)	127200492.5	1308447.423	155.5581	
人口 30 万以上市				0.000
30 万未満市, 町村				-18.857
雇用者数	125981712.7	89667.609	10.6604	-0.143

元の回帰モデルでは、４つの質的変数によって、産業小分類（12 区分）×経営組織（3 区分）×本所・支所の別（3 区分）×地域（3 区分）の 324 通りの組み合わせが作られる。カテゴリー統合後のモデルでは、産業小分類（6 区分）×経営組織（2 区分）×本所・支所の別（3 区分）×地域（2 区分）の 72 通りに減り、カテゴリー間の相違が明確になる。例えば、産業小分類では、「70A 一般食堂」と「70C 西洋料理店」、「70B 日本料理店」と「703 すし店」は、それぞれ（１雇用者当たり給与支給総額の点で）似ている。また、「70A 一般食堂、70C 西洋料理店」に対して、「70B 日本料理店、703 すし店」は約 20 万円高い水準であることなどが分かる。

この場合の重回帰式は、次のとおりである。説明変数として追加した産業小分類、経営

組織、本所・支所の別、地域は、雇用者数 x_i の係数として働く点に注意する。さらに、これらの因子別に x_i^2 の係数を変えるには、産業小分類のダミー変数×雇用者数などの項を説明変数として追加することで拡張できる。同様に、質的変数の交互作用項を入れて拡張することもできる。

$$\hat{y}_i = \left\{ 146.356 + \begin{pmatrix} 0.000 \\ 19.900 \\ 11.944 \\ -6.035 \\ -15.883 \\ -51.513 \end{pmatrix} \begin{matrix} \text{70A,70C} \\ \text{70B,703} \\ \text{70D} \\ \text{70E,702,704} \\ \text{70F,70J,70H} \\ \text{70G} \end{matrix} + \begin{pmatrix} 0.000 \\ 59.078 \end{pmatrix} \begin{matrix} \text{個人経営} \\ \text{会社、その他} \end{matrix} \right.$$

$$\left. + \begin{pmatrix} 0.000 \\ 70.848 \\ -40.409 \end{pmatrix} \begin{matrix} \text{単独} \\ \text{本所} \\ \text{支所} \end{matrix} + \begin{pmatrix} 0.000 \\ -18.857 \end{pmatrix} \begin{matrix} \text{30万以上市} \\ \text{30万未満市、町村} \end{matrix} - 0.143 \times x_i \right\} \times x_i$$

（3）2つの方法の比較

（1）、（2）では、層化されていない単回帰モデルの残差を多段層別分析によって分類する方法と、現行の補定法における層化変数を説明変数に入れた重回帰・数量化Ⅰ類モデルによって、層化因子の効果を確認する方法を、事例を挙げて説明した。同一のデータやモデルにより、さらなる比較分析が必要であるが、いくつかの事例から得られたことを以下に述べる。

現行の補定法と同様の一律の層化基準をとると、完全データの大きさが十分ではない層が多数生じる。多段層別分析では、逐次的に 2 分割を繰り返すことで、完全データの大きさを確保して、少ない数の層（ノード）を設定することができる。層化変数を説明変数に入れた重回帰・数量化Ⅰ類モデルの適用では、カテゴリーの類似性を考慮した層化の効果が確認できる。どちらも回帰モデルの構築と統合的な方法であるが、2つの方法で得られた層化が一致しない場合がある。業種業態などの背景領域の情報を取り入れるといった専門的な判断が必要となるだろう。

５．収入額の回帰補定

ここで、収入額の回帰補定について簡単に説明する。収入額の補定は、従業者数と収入額の相関関係を回帰モデルで利用する。3、4 節で検討した結果から、現行の補定法における層化変数を説明変数に入れた重回帰・数量化Ｉ類モデルを中心に検討した。また、説明変数の値の小さいところでは、前述したように、回帰補定の代わりに平均値補定を部分的に用いた。

中分類「69 不動産賃貸業・管理業」を例に取ると、経営組織が個人で従業者数が１人以下の事業所は、産業小分類×地域の層別に平均値で補定し、個人で従業者２人以上と会社の事業所は、従業者数、産業小分類、経営組織、本所・支所の別、地域を説明変数とする重回帰・数量化Ｉ類モデルを用いる。１従業者当たりのモデルと対数線形のモデルを相対予測残差で比較すると、対数線形が選択された。なお、他の産業についても、１従業者当たりのモデルよりも、対数線形のモデルが選択される傾向がある。

さらに、回帰補定を適用した後、収入額合計について公式統計書の数値と比較したところ、回帰補定による収入額は過小になる傾向が見られた。収入額階級別事業所数を用いて分布を比較すると、回帰補定は、特定の収入額階級に集中する傾向が確認できた。現行の補定法と比べて層の数が少ないため、同一の値に集まった結果であると考えられる。しかし、この補定値が過小であるかどうかは、「真値」と比較しなければならない。このような補定の評価に必要な「真値」をどのように決めるかという問題は、今後の課題である。

６．経済センサスの経理項目補定への応用

本稿では、平成 16 年サービス業基本調査の経理項目補定について、線形回帰モデルを拡張した対数線形回帰モデルと１人当たり回帰モデルの当てはめ、補定の観点からの相対予測残差によるモデルの評価、回帰モデルの構築と統合的に層を構築する試みを説明した。

まとめに代えて、今後の経済センサスの経理項目補定への応用に参考になると思われる点を以下に挙げる。

- 分散不均一への対応として、対数線形回帰モデルや原単位回帰モデルの導入を検討する
- 雇用者数・従業者数等の説明変数が小さいところについては、回帰補定の他に、平均値補定などの適用可能性を検証する必要がある
- 雇用者数・従業者数等の説明変数が非常に大きいところは、回帰式の補外の適用は慎重に行うべきで、結果精度への影響が大きいことから、できる限り欠測をなくすことが必要である
- 相対予測残差によるモデルの評価が有効であると思われるが、さらに詳細な性質を調べる必要がある
- 相対予測残差による外れ値の検出に効果が見られるが、最終的な判断には業種業態などの背景領域の知識が重要である

- 完全データの大きさが十分ではないときは、一律の層化・合併基準を適用するのではなく、回帰モデルと統合的でシンプルな基準を探索する方法を利用できるが、最終的な判断には背景領域の知識の活用が重要である
- 上記の方法に関して、回収率がより低くなった場合の影響を検討する必要がある
- 補定の評価のために、「真値」を得る方法を検討する必要がある

経済センサスの経理項目補定の流れとしては、まず、平成 16 年サービス業基本調査のデータや、先行処理した一部のデータで初期推定値を得て補定を行い、最終的には、チェック済みの全データを使った推定によって更新することが考えられる。しかし、平成 16 年データの使用は、産業によっては調査時点の違いが大きいことが予想される。さらに、従来のサービス業基本調査の対象外の産業の検討も必要になる。また、公表までの時間の制約から、初期推定、更新といった段階的な検討の時間が十分に取れないことも予想される。これらについては、試験調査の活用や、回帰モデル構築等に必要な多変量解析手法についての事前のトレーニングが有効と考えられる。

参考文献

統計センター事業企画課 (2005). 平成 16 年サービス業基本調査の経理項目の補定方法について（メモ），平成 17 年 6 月 7 日研究センター内会議資料.

Thompson, K.J. and Williams, Q. (2003). Developing Imputation Models for the Services Sectors Portion of the Economic Census, *Statistical Policy Working Paper 37*, US Federal Committee on Statistical Methodology.

Williams, Q. and Thompson, K.J. (2004). Evaluating Regression Imputation Models: An Example from the Services Sectors Portion of the Economic Census, *Proceedings of the Annual Meeting of the American Statistical Association 2004*.

製 表 技 術 参 考 資 料 8

平成 20 年 3 月発行

編集・発行　　独立行政法人　統計センター

〒162－8668
東京都新宿区若松町 19－1
電　話　代表　03 (5273) 1200

研究センター　TEL : 03－5273－1286
E-mail : research@nstac.go.jp