Monthly rent of condominiums in Tokyo

Giacomo De Leva

1 Introduction

Tokyo Metropolis, commonly referred as Tokyo, is one of the 47 prefectures in Japan and capital of the country. Tokyo Metropolis is divided in 23 special wards, which cover the area that was originally the City of Tokyo before it was abolished in 1943 to become part of the newly-created Tokyo Metropolis, and 39 municipalities in the western part of the prefecture and the two outlying island chains. The total population of the prefectures exceeds 13 million people, while the population in the special wards is over 8 million people. In this report we want to model the rent price of condominiums in the special wards.

Figure 1: The 23 special wards of Tokyo Metropolis (from Wikipedia)

2 Data

Data about rent prices in the 23 special wards were extracted by means of a script written in Perl from the website http://tokyo-athome.jp, which is one of the most popular Japanese sites for searching for condominiums or houses to rent or buy. The data extracted include information about 233,445 condominiums. More precisely, we have 233,445 items of the following 9 variables:

- Z_1: The monthly rent expressed in Yen (JPY).
- Z_2: Building maintenance fees (building management fees, electricity costs for common areas and cleaning costs) expressed in Yen.
- Z_3: The address of the condominium. This includes the ward (for example "Bunkyo"), the location within the ward (for example "Sendagi"), and the city district (for example "4 chome"), but not the city block and the house number.

- Z_4: Time in minutes to walk from the condominium to the nearest train or subway station (time is calculated on the basis that one minute is needed to walk 80 meters).
- Z_5: Number of the rooms of the condominium expressed by a code. Each code has one of the following formats: "1R", "nK", "nSK" ,"nDK" , "nLK", "nSDK", "nSLK", "nLDK", and "nSLDK". The code "1R" means one-roomed flat (kitchen inside the room). In the other codes n is an integer that denotes the number of rooms in the condominium and the combination of the letters "K", "S", "D", and "L" indicates which common areas there are, where "K", "S", "D" and "L" mean "Kitchen", "Storage room" (usually, a Walk-in closet), "Dining room ", and "Living room" respectively. For example "5LK" means that the condominium has five rooms, a kitchen, and a living room.
- Z_6: The surface of the condominium expressed in square meters.
- Z_7: Building type of the condominium. There are two types: "Apaato", often older buildings, which are usually only a few stories in height, without a central secure entrance, and "Manshon" more modern expensive buildings with multiple floors, elevators, and a communal secure gate. Buildings of this type are usually more sturdily built than those of "apaato" type, normally of reinforced concrete construction. Though commonly accepted standards for description exist, this is not a legal requirement, therefore descriptions may not be entirely accurate (from Wikipedia).
- Z_8: Built year of the condominium.
- Z_9: Two numbers that indicate respectively the number of stories of the building and the story of the condominium.

Since the monthly payment is given by the sum of the rent and the maintenance fees, we define the response variable Y as $Y := Z_1 + Z_2$. We include in our analysis also the variables Z_6 and Z_4 that we rename to X_1 and X_2, respectively. Instead of considering the built year of the condominium, we prefer to consider its age in years. Hence, we define $X_3 := 2013 - Z_8$. The variable Z_3 has 2968 different values. In order to simplify, we define two variables X_4 which includes the ward and X_5 which includes the location within the ward. Since we do not expect sensible changes of the response variable within the same district, we will not consider the city district in our analysis. We rename the variable Z_7 to X_6. From the variable Z_9 we extract only the number of stories of the building and define the variable X_7 by that value. We delete 420 items because they have at least one missing value in one of the variables X_1-X_7. Finally, we consider the variable Z_5. This variable assumes 54 values, but many values are quite rare. For example there is respectively only one item for "8DK" and "8LDK". We decided to consider only the most 21 frequent values and to delete the remaining 443 items. We rename Z_5 to X_8. Our data set for analysis includes 232,582 items.

Variable	Value
Y	Rent and fees
X_1	Surface of the condominium
X_2	Time distance from nearest station
X_3	Age of the building
X_4	Ward
X_5	Location
X_6	Building type
X_7	Floor
X_8	Number of rooms

Variables for the Tokyo condominium data

The variable X_8 may assume one of the following values: 1R, 1DK, 1K, 1LDK, 1LK, 1SDK, 1SK, 1SLDK, 2DK, 2K, 2LDK, 2SDK, 2SLDK, 3DK, 3K, 3LDK, 3SDK, 3SLDK, 4DK, 4LDK, 4SLDK. In order to look for corrupted items we produce a simple scatterplot of Y against X_1 (see Figure 2).

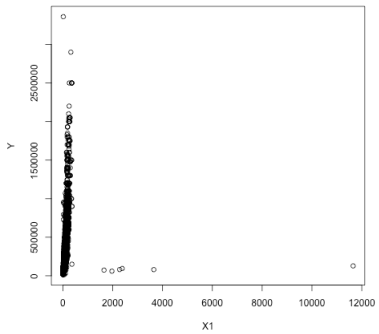


Figure 2: Scatterplot of Y against X_1

We observe the presence of 8 corrupted data. Indeed, one condominium has a rent greater than 3,000,000 (about \$30,000) while the surface is only 16 m^2. The other 7 items have a too large surface to be plausible. We delete this items. After deleting these items we produce again a scatterplot of Y against X_1 (see Figure 3).

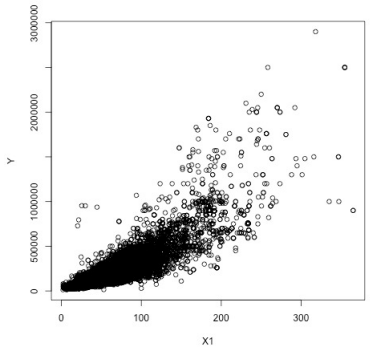


Figure 3: Scatterplot of Y against X_1 after deleting corrupted items

Investigating the values of the variables X_2, we notice that there are values for this variable bigger than 35, equivalent to 2.8 km to the nearest station. This value is not plausible in the 23 wards where . Hence, we decide to delete these 83 items. Figure 4 shows parallel boxplots for Y by X_2. As expected the values of the rent by each value of the distance to the nearest train station have a large spread, but the mean values indicate that in general the value of the rent decreases by increasing the distance to the nearest station.

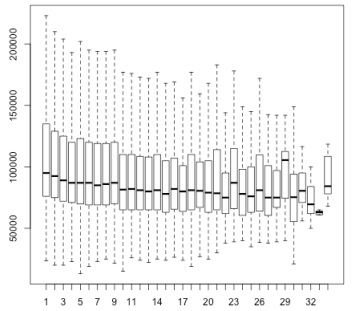


Figure 4: Paraller boxplots for Y by X_2

We also notice that there are 88 condominiums that belong to buildings built before the end Second World War corresponding to values of the variable X_3 larger than 68. These values may be plausible,

but nevertheless we prefer to exclude them from our analysis. Figure 5 shows parallel boxplots for Y by X_3. As before the values of the rent by each value of the age of the building have a large spread, but the mean values indicate that in general the value of the rent decreases by increasing the age.

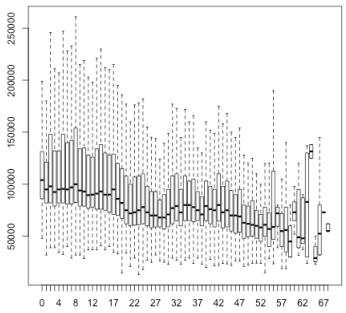


Figure 5: Paraller boxplots for Y by X_3

Figure 6 shows parallel boxplots for Y by the variables X_4, X_6, X_7, and X_8, respectively.

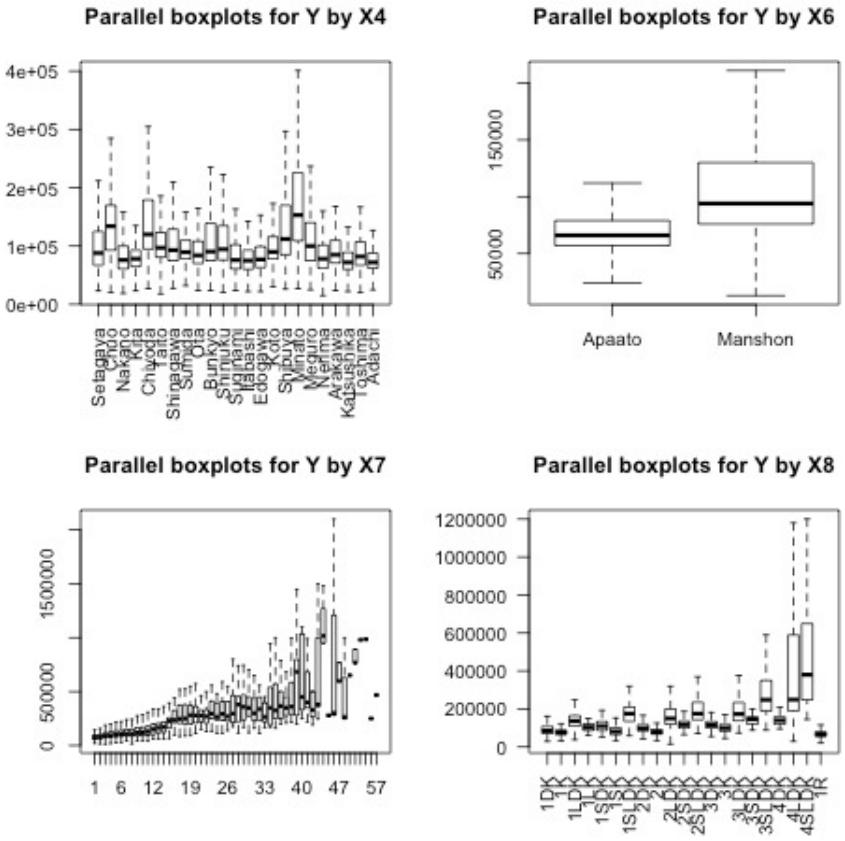


Figure 6: Paraller boxplots for Y by X_4, X_6, X_7, and X_8

Finally, we remark that the variable X_5 assumes 862 values.

3 Modeling

One of the most important assumptions of the linear regression model is the normality of the response variable. Applying the one-sample Kolmogorov-Smirnov test to the whole sample of Y, we find out that

we can reject this hypothesis with a p-value smaller than 10^{-16}. We have the same conclusion even if we restrict the sample of Y to a particular ward. Hence, we can model rent prices neither with a single linear model valid for all 23 wards, nor with a model valid for a particular ward. We need to consider models valid for a particular location, that is for a fixed value of the variable X_5. We only consider locations for which there are at least 11 items and the p-value of the Kolmogorov-Smirnov test for the sample Y restricted to that location is at least 0.1 (for locations with a smaller p-value it may be necessary to restrict the sample to the level of district, bur we shall not perform this analysis). The number of locations that satisfy this criteria is equal to 97. We shall focus our attention to the location Wakamiyacho of the Shinjuku ward. We chosen this location because the rent prices on this location has the maximal variance and one of the smallest p-value in the Kolmogorov-Smirnov test (0.102). We believe that rent prices in this location should be the most difficult to fit in a linear regression model.

Y	X_1	X_2	X_3	X_6	X_7	X_8
78,000	23.68	2	8	Apaato	2	1K
75,000	14.27	5	23	Manshon	4	1R
84,000	21.00	5	25	Manshon	1	1R
95,000	33.00	6	45	Manshon	3	1K
119,000	31.72	3	23	Manshon	5	1LDK
119,000	31.72	4	23	Manshon	5	1LDK
115,000	36.00	8	45	Manshon	3	2DK
144,000	42.83	5	25	Manshon	1	1DK
144,000	42.08	6	25	Manshon	1	2K
247,000	51.79	3	4	Manshon	2	3K
650,000	164.16	5	13	Manshon	1	3LDK
750,000	215.19	5	13	Manshon	2	4LDK
1,000,000	235.85	7	6	Manshon	3	4LDK
1,050,000	251.59	7	6	Manshon	3	4SLDK
1,100,000	252.64	5	13	Manshon	2	4LDK
1,500,000	315.81	5	13	Manshon	4	4LDK

Data for the Wakamiyacho location

We shall use a forward selection approach, that is we start to fit the response variable in a linear model with the explanatory variable which is most correlated to the response variable and then we test if the addition of one more explanatory variable improves the model by means of the Akaike information criterion (AIC). In positive case we integrate the variable in the model. We repeat this procedure until the addition of one more variable does not longer improve the model. We decide a priori not to use the variable X_8 since we do not have enough data to take in account all the possible values of this variable. Among the explanatory variables, X_1 is the most correlated variable to Y. Hence, we consider the following model $M1$:

$$Y = a_1 X_1 + b + \epsilon.$$

```
Call:
lm(formula = Y ~ X1, data = wakamiyacho_data)

Residuals:
    Min      1Q  Median      3Q     Max
-166621  -15132   -1486   19467  140340

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -30882.3    23078.8  -1.338    0.202
X1            4403.1      152.3  28.920 6.92e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 63380 on 14 degrees of freedom
Multiple R-squared: 0.9835,Adjusted R-squared: 0.9824
F-statistic: 836.3 on 1 and 14 DF,  p-value: 6.919e-14
```

We notice that the adjusted R-squared value is very high, but the residual standard error as well (ca. \$630). Moreover, the p-value of the intercept suggests that maybe we should consider the following model $M2$:

$$Y = a_1 X_1 + \epsilon.$$

```
Call:
lm(formula = Y ~ -1 + X1, data = wakamiyacho_data)

Residuals:
    Min      1Q  Median      3Q     Max
-165628  -38194  -18238     919  156237

Coefficients:
   Estimate Std. Error t value Pr(>|t|)
X1   4255.0      107.3   39.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 65030 on 15 degrees of freedom
Multiple R-squared: 0.9906,Adjusted R-squared: 0.9899
F-statistic:  1574 on 1 and 15 DF,  p-value: < 2.2e-16
```

Model $M2$ suffers also from a very high residual standard error. We try the following model $M3$:

$$\log Y = a_1 \log X_1 + b + \epsilon.$$

```
Call:
lm(formula = log(Y) ~ log(X1), data = wakamiyacho_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26669 -0.11945 -0.00029  0.06884  0.36908

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.09081    0.17245   46.92  < 2e-16 ***
log(X1)      1.04033    0.03996   26.03 2.94e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*\ 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1646 on 14 degrees of freedom
Multiple R-squared: 0.9798,Adjusted R-squared: 0.9783
F-statistic: 677.8 on 1 and 14 DF,  p-value: 2.938e-13
```

Model $M3$ seems to be better than models $M1$ and $M2$. We try to integrate a second explanatory variable in model $M3$. The only variable who leads to a smaller AIC value with respect to the model $M3$ is the variable $\log X_3$. Hence, we consider the model $M4$ defined below:

$$\log Y = a_1 \log X_1 + a_3 \log X_3 + b + \epsilon.$$

```
Call:
lm(formula = log(Y) ~ log(X1) + log(X3), data = wakamiyacho_data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.24242 -0.07023 -0.00891  0.05622  0.35151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.59821    0.31274  27.493 6.64e-13 ***
log(X1)      0.99844    0.04297  23.236 5.68e-12 ***
log(X3)     -0.12071    0.06412  -1.883   0.0823 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1514 on 13 degrees of freedom
Multiple R-squared: 0.9841,Adjusted R-squared: 0.9817
F-statistic: 402.3 on 2 and 13 DF,  p-value: 2.039e-12
```

We try to add another variable to the model $M4$. The only variable that improves model $M4$ is the categorical variable X_6:

$$\log Y = a_1 \log X_1 + a_3 \log X_3 + X_6 + b + \epsilon. \tag{1}$$

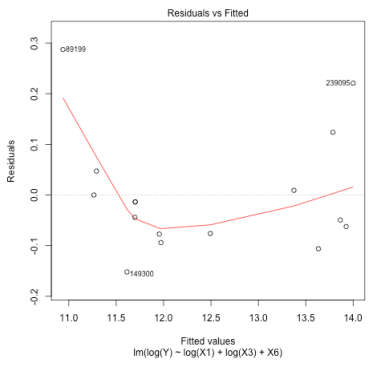
```
Call:
lm(formula = log(Y) ~ log(X1) + log(X3) + X6, data = wakamiyacho_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15199 -0.07629 -0.02879  0.01870  0.28728

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.62966    0.27521  31.356 6.97e-13 ***
log(X1)      0.95456    0.04271  22.348 3.80e-11 ***
log(X3)     -0.18563    0.06362  -2.918   0.0129 *
X6Manshon    0.35300    0.16058   2.198   0.0483 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1331 on 12 degrees of freedom
Multiple R-squared: 0.9887,Adjusted R-squared: 0.9858
F-statistic: 348.8 on 3 and 12 DF,  p-value: 6.196e-12
```

After that we notice that adding one more explanatory variable does not improve the model which is given in its definitive form by (1).



While the residual standard error of the model is relative small, the plot of residuals against fitted values shows three points that are very large in absolute value. Moreover, we cannot be so sure that there are not any pattern in the residuals.

4 Conclusions

We have seen that it is not possible to fit a linear model for the response variable Y valid for all the 23 wards of Tokyo. It is necessary to restrict the sample of data to a particular location corresponding to a fixed value of the variable X_5. As an example how such an analysis can be carried out we considered the data sample of the location Wakamiyacho. We have seen that rent prices in this location are affected in decreasing order of importance by the following variables: Surface (X_1), Age of the Buiding (X_3), and Type of the Building (X_6). However, we do not expect that the same conclusion still holds for other locations. In particular, we may expect that under some conditions also the floor number (X_7) may affect rent prices in a sensible way in locations where the range of the variable X_7 is broader.

Appendix

Below we report the Perl script used to extract the data from the website http://tokyo-athome.jp.

```
use strict;
use LWP::UserAgent;
my $ua = new LWP::UserAgent;
open (MYFILE, '>data.txt');
print MYFILE "Rent\tFees\tAddress\tWalk\tType\tSurface\tBuilding\tYear\Floor\n";
foreach my $page (1..2393) {
my $link = "http://tokyo-athome.jp/ajax/list/list?ART=01&CHIKUNENSU=kn001&DISPID=PB0H
            &DOWN=1&EKITOHO=ke001&ITEM=kr&ITEMNUM=100&JOHOKOKAI=kj001&MENSEKI=kt001&P
            AGENO=$page&PRICEFROM=kc001&PRICETO=kc138&SHIKU=13101%2C13102%2C13103%2C1
            3104%2C13105%2C13106%2C13107%2C13108%2C13109%2C13110%2C13111%2C13112%2C13
            113%2C13114%2C13115%2C13116%2C13117%2C13118%2C13119%2C13120%2C13121%2C131
            22%2C13123&SITECD=40013&TJOKENCD=kb001%2Ckb002%2Ckb003%2Ckc001%2Ckc138%2C
            kc201%2Ckc202%2Ckc203%2Ckm002%2Ckm003%2Ckm004%2Ckm005%2Ckm008%2Ckm009%2Ck
            m010%2Ckm013%2Ckm014%2Ckm015%2Ckm018%2Ckm019%2Ckm021%2Ckt001%2Cke001%2Ckn
            001%2Ckj001%2Ckj002%2Ckj003%2Ckj004%2Cka001%2Ckg001%2Ckg002%2Ckg008%2CB01
            %2CA01%2CP02%2CO04%2CO20";
my $response = $ua->post($link);
my $content = $response->content;
my @infos = split(/<p class="price">/, $content);
foreach my $info (@infos) {
         my $rent;
         my $fees;
         my $address;
         my $walk;
         my $type;
         my $surface;
         my $building;
         my $year;
         my $floor;
         if ($info =` m/F<strong>([^<]*)</) {
             $rent = $1;
         }
         else {next}
         if ($info =` m/F<\/span>([^<]*)</) {
```

```
            $fees = $1;
        }
        if ($info =‘ m/p class="addr">\s+([^<]*)</) {
            $address = $1;
        }
        if ($info =‘ m/k\s*(\d*)/) {
            $walk = $1;
        }
        if ($info =‘ m/<\/th>\s*<td>([^<]*)<\/td>\s*<th scope="row"><\/th>
\s*<td>\s*([^\sm]*)/) {
            $type = $1;
            $surface = $2;
        }
        if ($info =‘ m/<\/th>\s*<td>\s*([^<]*)</) {
            $building = $1;
        }
        if ($info =‘ m/zN<\/th>\s*<td>\s*(\d+)/) {
            $year = $1;
        }
        if ($info =‘ m/K\/K<\/th>\s*<td>\s*([^<]*)</) {
            $floor = $1;
        }
        print MYFILE "$rent\t$fees\t$address\t$walk\t$type\t$surface\t$building\t$ye
ar\t$floor\n";
}
}
close (MYFILE);
```