Lecture 11: Linear regression analysis

IST5573

統計方法 Statistical methods

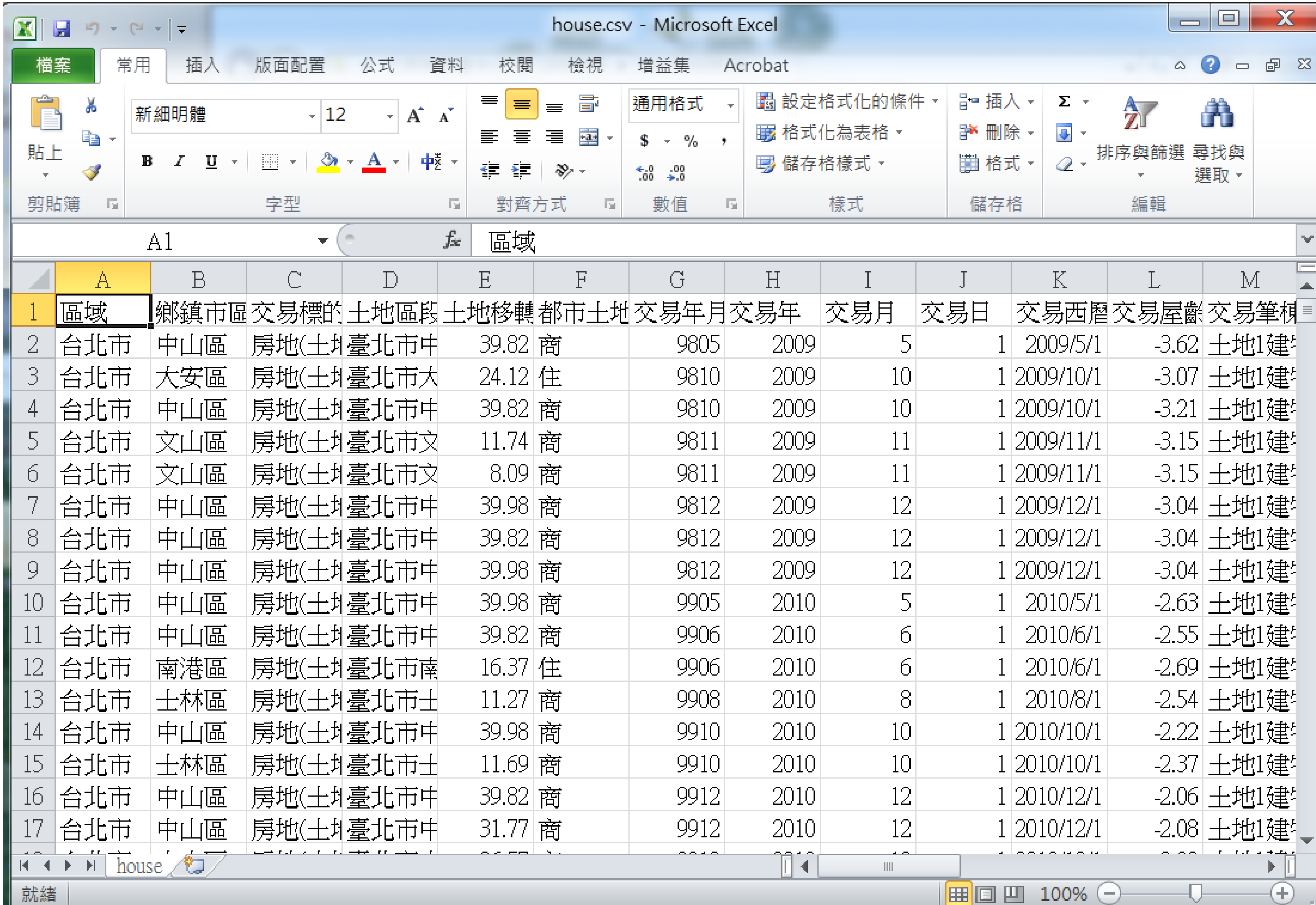
2016/11/23

Regression

- Regression is a statistical technique used to predict the value of response variable (Y) (or dependent variable) according to one or more covariate(s) (X) (or independent variable(s)).
- If the respond variable (Y) is continuous (e.g., weight, blood pressure), the linear regression model is used.
- If the respond variable (Y) is binary (e.g., success or failure), the logistic regression model is used.

政府資料開放平臺
http://data.gov.tw/

六都房地產實價登錄資料



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	區域	鄉鎮市區	交易標的	土地區段	土地移轉	都市土地	交易年月	交易年	交易月	交易日	交易西曆	交易屋齡	交易筆棟
2	台北市	中山區	房地(土地	臺北市中	39.82	商	9805	2009	5	1	2009/5/1	-3.62	土地1建物
3	台北市	大安區	房地(土地	臺北市大	24.12	住	9810	2009	10	1	2009/10/1	-3.07	土地1建物
4	台北市	中山區	房地(土地	臺北市中	39.82	商	9810	2009	10	1	2009/10/1	-3.21	土地1建物
5	台北市	文山區	房地(土地	臺北市文	11.74	商	9811	2009	11	1	2009/11/1	-3.15	土地1建物
6	台北市	文山區	房地(土地	臺北市文	8.09	商	9811	2009	11	1	2009/11/1	-3.15	土地1建物
7	台北市	中山區	房地(土地	臺北市中	39.98	商	9812	2009	12	1	2009/12/1	-3.04	土地1建物
8	台北市	中山區	房地(土地	臺北市中	39.82	商	9812	2009	12	1	2009/12/1	-3.04	土地1建物
9	台北市	中山區	房地(土地	臺北市中	39.98	商	9812	2009	12	1	2009/12/1	-3.04	土地1建物
10	台北市	中山區	房地(土地	臺北市中	39.98	商	9905	2010	5	1	2010/5/1	-2.63	土地1建物
11	台北市	中山區	房地(土地	臺北市中	39.82	商	9906	2010	6	1	2010/6/1	-2.55	土地1建物
12	台北市	南港區	房地(土地	臺北市南	16.37	住	9906	2010	6	1	2010/6/1	-2.69	土地1建物
13	台北市	士林區	房地(土地	臺北市士	11.27	商	9908	2010	8	1	2010/8/1	-2.54	土地1建物
14	台北市	中山區	房地(土地	臺北市中	39.98	商	9910	2010	10	1	2010/10/1	-2.22	土地1建物
15	台北市	士林區	房地(土地	臺北市士	11.69	商	9910	2010	10	1	2010/10/1	-2.37	土地1建物
16	台北市	中山區	房地(土地	臺北市中	39.82	商	9912	2010	12	1	2010/12/1	-2.06	土地1建物
17	台北市	中山區	房地(土地	臺北市中	31.77	商	9912	2010	12	1	2010/12/1	-2.08	土地1建物

Variable	Description
每平方公尺單價	元
豪宅	0=每平方公尺單價 ≤ 20萬 1=每平方公尺單價 > 20萬
區域	台北市、新北市、桃園市、台中市、台南市、高雄市
車位	0=無, 1=有
屋齡	建築完成到2015/9/18 (年)
主要用途	工業用、住家用、住商用、商業用、國民住宅
建物型態	公寓(5樓含以下無電梯)、住宅大樓(11層含以上有電梯)、店面(店鋪)、套房(1房1廳1衛)、透天厝、華廈(10層含以下有電梯)、廠辦、辦公商業大樓
有無管理組織	0=無, 1=有

Linear regression

Simple linear regression

- The respond variable (Y) follows a normal distribution
- Only one covariate X
- $$\mathrm{E}(Y) = \beta_0 + \beta_1 X$$

  Y: response variable (continuous) (known)

  X: covariate (continuous or binary) (known)

  β_0, β_1: regression coefficients (unknown)

Interpretation of regression coefficients

- $$\mathrm{E}(Y) = \beta_0 + \beta_1 X$$

  $\beta_0 =$ the average of Y when $X = 0$

  $\beta_1 =$ the average change of Y for every 1 unit increase in X

Example 1

- From六都房地產實價登錄資料:

  $E(\text{每平方公尺單價}) = 74862.95 - 3.52 \times \text{屋齡}$

  $\beta_0 = 74862.95$ = 屋齡為0時，房屋每平方公尺的平均單價

  $\beta_1 = -3.52$ = 屋齡每增加1年，每平方公尺平均單價將減少3.52元

- The average change in Y is the same for every 1 unit change in X, no matter what the value of X is (linearity).

(RMD_example 11.2)

Example 2

- E(每平方公尺單價) = 77456.9 − 2135.3 × 車位
- μ_Y = 有車位的房屋，其每平方公尺的平均單價

  μ_N =沒有車位的房屋，其每平方公尺的平均單價
- $\beta_0 = \mu_N =$ 77456.9 =沒有車位的房屋有，其每平方公尺的平均單價

  $\beta_1 = \mu_Y - \mu_N =$ -2135.3 = 有車位房屋和沒有車位的房屋，他們每平方公尺平均單價的差異

(RMD_example 11.2)

Parameter estimation: the least-squares method

- $\hat{y}_i =$ estimated response at x_i based on the fitted regression line
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated intercept and slope.
- Use the least-squares method to determine the best-fitting straight line (regression line):
choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize
$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Is there a significant linear relationship between y and x

- Use t-test or CI for β_1:

$$\text{Ho}: \beta_1 = 0 \quad \text{Ha}: \beta_1 \neq 0$$

How good the regression model is

- Coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{\text{variation due to regression}}{\text{total variation}}$$

- R^2 gives the proportion of total variability explained by regression.
- The larger the value of R^2, the better the fit of the regression model.

Regression results

$\hat{\beta}_0$

$\hat{\beta}_1$

```
Call:
lm(formula = 每平方公尺單價 ~ 屋齡)

Residuals:
   Min     1Q Median     3Q    Max 
-74775 -38131 -19608  19483 839722 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 74862.948   1104.675   67.77   <2e-16 ***
屋齡           -3.524     50.031   -0.07    0.944    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59790 on 10028 degrees of freedom
  (289 observations deleted due to missingness)
Multiple R-squared:  4.948e-07, Adjusted R-squared:  -9.923e-05 
F-statistic: 0.004962 on 1 and 10028 DF,  p-value: 0.9438
```

(RMD_example 11.2)

Regression results

$SE(\hat{\beta}_0)$

$SE(\hat{\beta}_1)$

```
Call:
lm(formula = 每平方公尺單價 ~ 屋齡)

Residuals:
   Min     1Q Median     3Q    Max
-74775 -38131 -19608  19483 839722

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 74862.948   1104.675   67.77   <2e-16 ***
屋齡           -3.524     50.031   -0.07    0.944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59790 on 10028 degrees of freedom
  (289 observations deleted due to missingness)
Multiple R-squared:  4.948e-07, Adjusted R-squared:  -9.923e-05
F-statistic: 0.004962 on 1 and 10028 DF,  p-value: 0.9438
```

(RMD_example 11.2)

Regression results

p−value for Ho: $\beta_0 = 0$

p−value for Ho: $\beta_1 = 0$

```
Call:
lm(formula = 每平方公尺單價 ~ 屋齡)

Residuals:
   Min     1Q Median     3Q    Max
-74775 -38131 -19608  19483 839722

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 74862.948   1104.675   67.77   <2e-16 ***
屋齡           -3.524     50.031   -0.07    0.944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59790 on 10028 degrees of freedom
  (289 observations deleted due to missingness)
Multiple R-squared:  4.948e-07, Adjusted R-squared:  -9.923e-05
F-statistic: 0.004962 on 1 and 10028 DF,  p-value: 0.9438
```

(RMD_example 11.2)

Regression results

R^2

Adj R^2

```
Call:
lm(formula = 每平方公尺單價 ~ 屋齡)

Residuals:
   Min     1Q Median     3Q    Max
-74775 -38131 -19608  19483 839722

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 74862.948   1104.675   67.77   <2e-16 ***
屋齡           -3.524     50.031   -0.07    0.944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59790 on 10028 degrees of freedom
  (289 observations deleted due to missingness)
Multiple R-squared:  4.948e-07, Adjusted R-squared:  -9.923e-05
F-statistic: 0.004962 on 1 and 10028 DF,  p-value: 0.9438
```

(RMD_example 11.2)

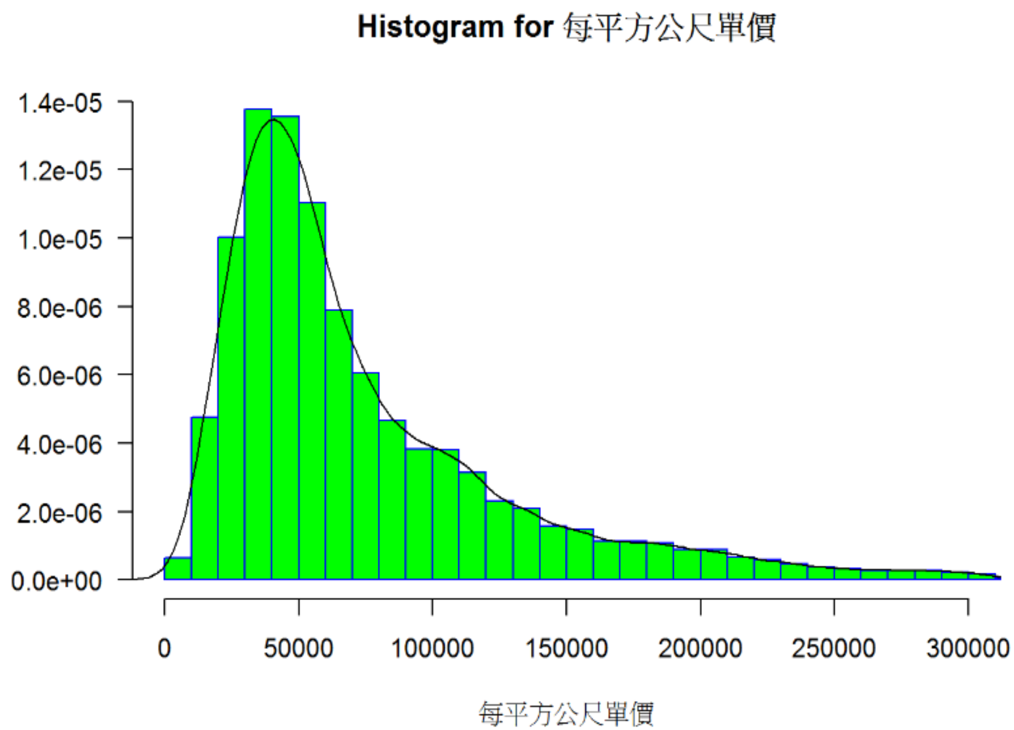
The residual plot

- Important to check the assumptions of a regression analysis (model diagnosis).
- It is most straightforward by viewing residuals

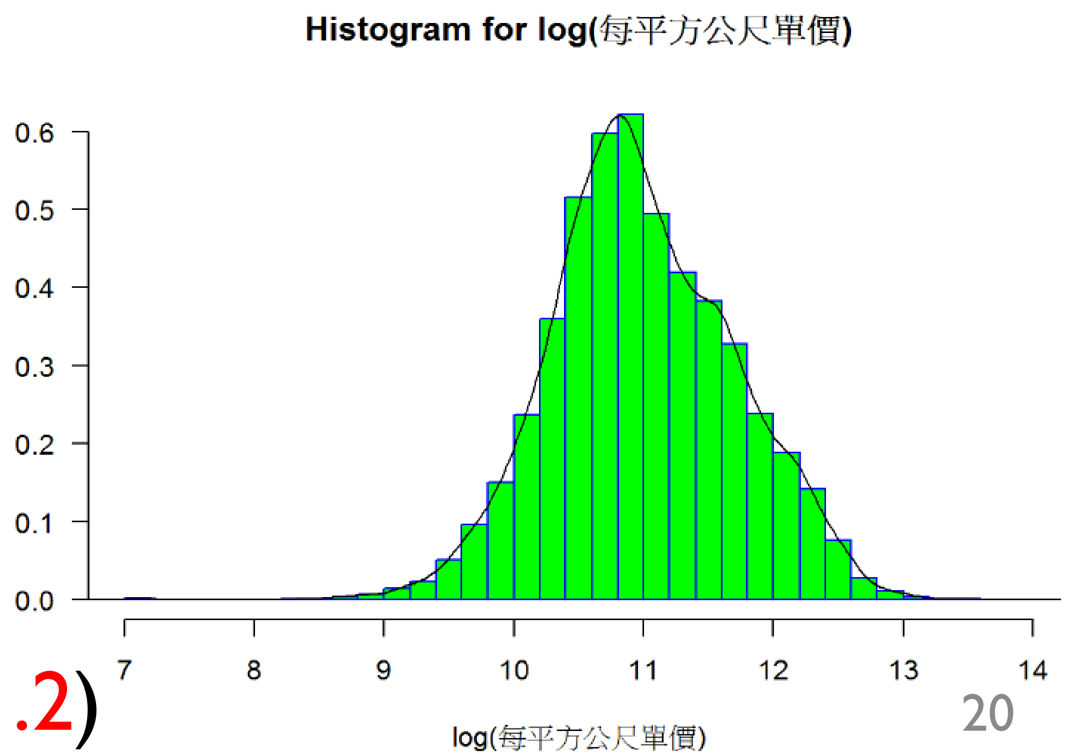
$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- Residuals are computed for each observation, and are usually plotted in at least two ways:
    - A scatter plot of e_i versus the predicted values $\hat{y}_i$ or the independent variable x_i.
    - “No pattern” -> good fit

- Draw histogram or q-q plot on y_i's or e_i's to check the normality
  - Data are skewed.
    1. If right skewed, transfer y to $\sqrt{y}$ or $\ln(y)$.
    2. If left skewed, transfer y to y^2 or e^y.
- Outlier: a set of residuals is much larger than the rest in absolute value, perhaps, lying three or more standard deviations from the mean of the residuals.

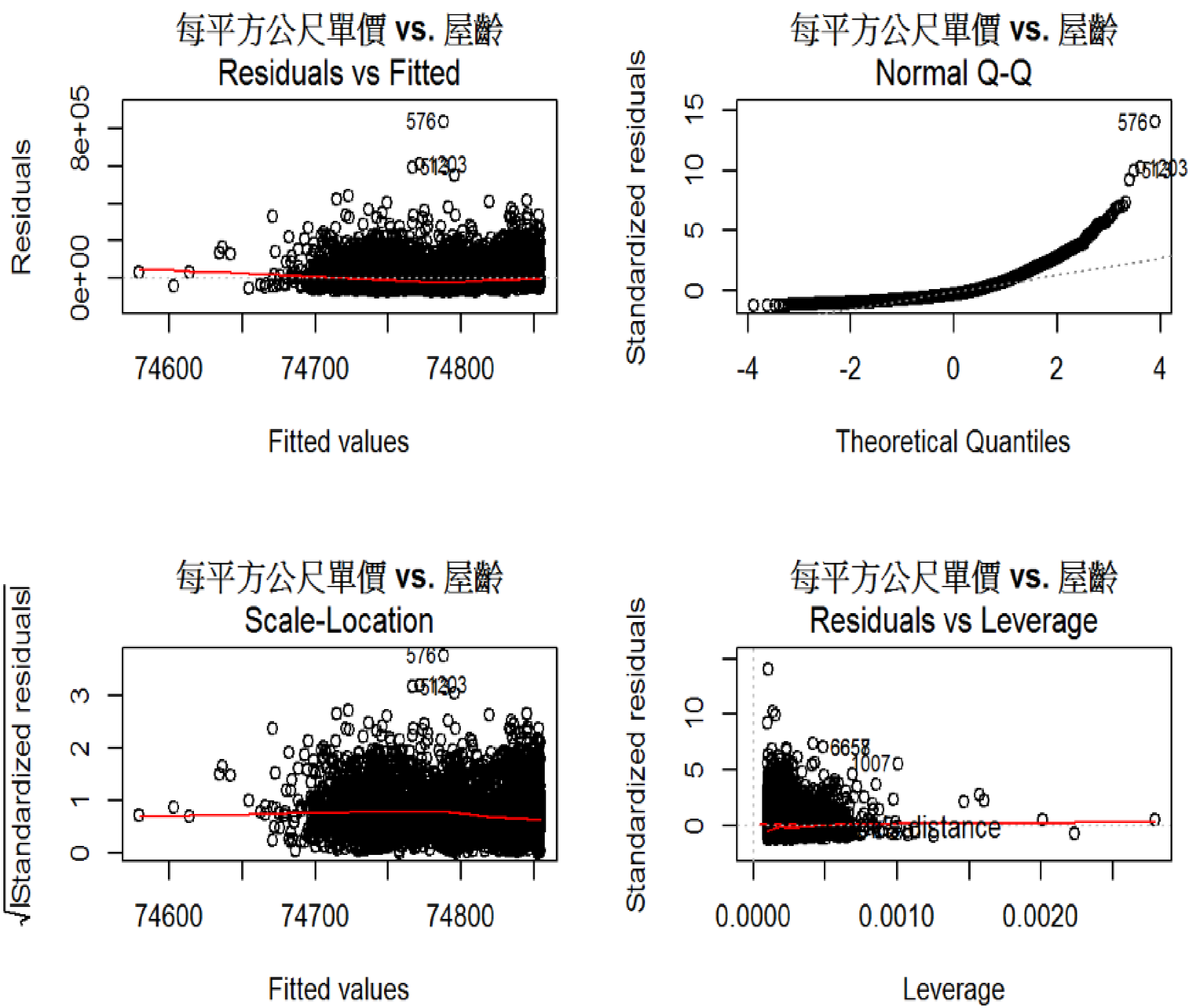


Display a smooth density estimate



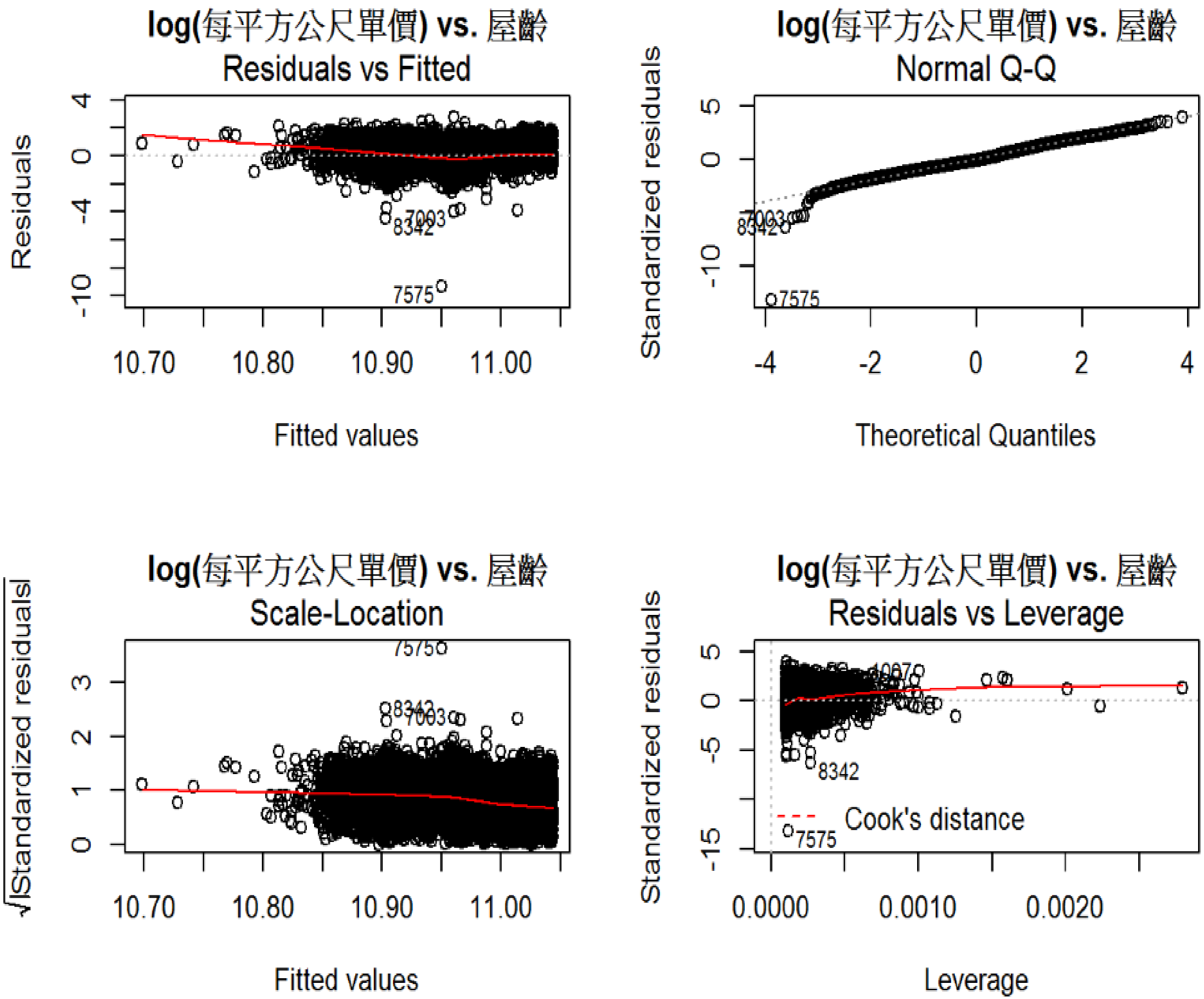
(RMD_example 11.2)

The residual plot on y



(RMD_example 11.2)

The residual plot on log(y)



(RMD_example 11.2)

Multiple linear regression

- When several X's are used as covariates, we have multiple linear regression.
- $$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$$
- Each coefficient describes the linear relationship between Y and X controlling (adjusting) for all the other X's (or, in other words, holding the other X's constant).

Example

- $E(\text{每平方公尺單價}) = 74117.99 + 963.11 \times \text{車位} + 17.18 \times \text{屋齡}$
- $\beta_1 = 963.11$ = 對那些屋齡相同的房屋，有車位和沒有車位房子，他們每平方公尺平均單價的差異
- Here, we assume that the relationship between "每平方公尺單價" and "車位" is the same at all "屋齡".
  - This is the parallelism assumption, or no interaction.

(RMD_example 11.3)

Model fit in multiple linear regression

- R^2 increases when additional covariates are added to the model.
- Adjusted coefficient of determination

$$\text{Adj } R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-K-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}$$

- Adjusted R^2 takes the number of covariates into account, and is useful when comparing models with different numbers of covariates.

Polynomial regression

- When the relationship between Y and X is nonlinear
- $\mathrm{E}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots$
- There may also be covariates besides X.

Example

- $E(\text{每平方公尺單價}) = 93179.05 - 2674.87 \times \text{屋齡} + 64.23 \times \text{屋齡}^2$
- The change of "每平方公尺單價" for "屋齡" increasing from 10 to 11 is different from the change for "屋齡" increasing from 30 to 31

(RMD_example 11.4)

Dummy variables

- Often in the situation where we want to compare more than two groups.
- Let $X = 1, 2, \cdots, M$ represent different groups. We can enter $X, X^2, X^3, \cdots$ into a regression equation if we are interested in modeling the "trend".
- If we are more interested in estimating individual differences between the groups, this situation calls for the use of dummy variables.

How to create dummy variables

- To compare several (say M) groups:
  1. Choose a "baseline group" with which to compare all others.
  2. There are $M - 1$ possible comparisons with a baseline group, so we need $M - 1$ dummy variables.

Example

- "區域" groups：台北市、新北市、桃園市、台中市、台南市、高雄市
  - $X_{北} = \begin{cases} 1 \text{ if 區域} = \text{台北市} \\ 0 \text{ otherwise} \end{cases}$ $X_{新} = \begin{cases} 1 \text{ if 區域} = \text{新北市} \\ 0 \text{ therwise} \end{cases}$

    $X_{桃} = \begin{cases} 1 \text{ if 區域} = \text{桃園市} \\ 0 \text{ otherwise} \end{cases}$ $X_{中} = \begin{cases} 1 \text{ if 區域} = \text{台中市} \\ 0 \text{ therwise} \end{cases}$

    $X_{南} = \begin{cases} 1 \text{ if 區域} = \text{台南市} \\ 0 \text{ therwise} \end{cases}$
  - "區域" = 高雄市 as the baseline group, and 5 dummy variables: $X_{北}, X_{新}, X_{桃}, X_{中}, X_{南}$

(RMD_example 11.5)

Example

- E(每平方公尺單價) = $43377 + 137581\, X_{北} + 44628\, X_{新} + 5079\, X_{桃} + 6035\, X_{中} - 10489\, X_{南}$
- 43377 = 高雄市每平方公尺的平均單價

  137581 = 台北市與高雄市每平方公尺平均單價的差異

  44628 = 新北市與高雄市每平方公尺平均單價的差異

  5079 = 桃園市與高雄市每平方公尺平均單價的差異

  …

(RMD_example 11.5)

Interaction in regression

- Interaction means that the association between the response Y and a covariate X_1 depends on the level of another covariate X_2.
- $$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$
- When interaction, the parallelism assumption is not true.

Example

- If the relationship between "每平方公尺單價" and "車位" is not the same for different "屋齡分組":

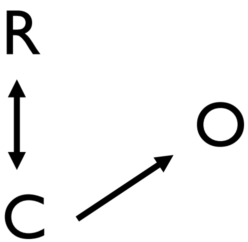
$$屋齡分組 = \begin{cases} 1, & 屋齡 > 25 \\ 0, & 屋齡 \leq 25 \end{cases}$$

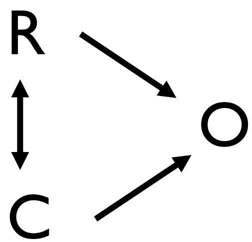
- E(每平方公尺單價) = 66790.6 + 8011.5 (車位) + 18276.6 (屋齡分組) + 12951.1 (車位 × 屋齡分組)
  - 8011.5 = 對那些屋齡小於等於25年的房屋，有車位和沒有車位房子，他們每平方公尺平均單價的差異
  - 8011.5+12951.1 = 對那些屋齡大於25年的房屋，有車位和沒有車位房子，他們每平方公尺平均單價的差異

(RMD_example 11.6)

Confounding

POTENTIAL CONFOUNDER:





→ causal

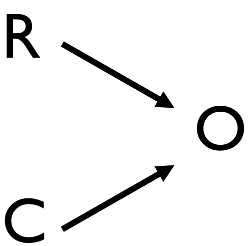
↔ associated

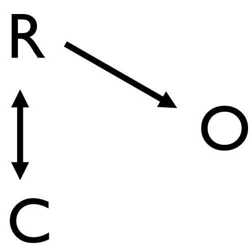
R=車位 (risk)

O=每平方公尺單價 (outcome)

C=屋齡分組 (confounder)

NOT A POTENTIAL CONFOUNDER:





Confounding in regression

- $\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  $\mathrm{E}(Y) = \beta^*{}_0 + \beta^*{}_1 X_1$
- Confounding if β_1 is very different from $\beta^*{}_1$
- The association between Y and X_1 changes substantially when X_2 (confounder) is included in the model.
- When confounding occurs and we are interested in associating Y with X_1, it is appropriate to adjust for X_2 (i.e., include X_2 in the model).

Example

- E(每平方公尺單價) = 66603.4 + 8382.0 (車位) + 18719.1 (屋齡分組)
- E(每平方公尺單價) = 77456.9 − 2135.3 (車位)
- $8382.0 \neq -2135.3$, "屋齡分組" is a confounding effect of the association between "每平方公尺單價" and "車位"

(RMD_example 11.7)

Variable selection

- Two "conflicting" goals in regression model building:
  1. Want as many covariates as possible so that the "information content" in the variables will influence $\hat{y}$.
  2. Want as few covariates as necessary because the variance of $\hat{y}$ will increase as the number of covariates increases.
- A compromise between the two hopefully leads to the best regression equation.

Criteria for evaluating subset regression models

Consider regression model:

$$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$$

1. Adjusted coefficient of determination:

$$\text{Adj } R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-K-1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)}$$

- This value will not necessarily increase as additional terms are introduced into the model. We want a model with the maximum $\text{Adj } R^2$.

2. Akaike information criterion (AIC) and Bayesian information criterion (BIC):

$$\mathrm{AIC} = -2\ln(L) + 2(K+1)$$

$$\mathrm{BIC} = -2\ln(L) + (K+1)\ln(n)$$

where L is the likelihood (the probability of observing our responses $y_1, \cdots, y_n$)

- AIC and BIC are log-likelihood measures penalizing the number of covariates in the model. BIC places a greater penalty on adding covariates as the sample size increases.
- Models with small values of AIC or BIC are preferred.

Variable selection procedure: all possible regressions

- If there are K covariates, we would investigate 2^K possible regression equations.
- Use the criteria above to determine some candidate models and complete regression analysis on them.
- R package leaps performs an all possible regressions methodology.

(RMD_example 11.8)

Stepwise regression methods

- Three types of stepwise regression methods:
  1. backward elimination
  2. forward selection
  3. stepwise regression (combination of forward and backward)

Backward elimination

1. Starting with all candidate covariates
2. Testing the deletion of each covariate using a chosen model fit criterion, deleting the covariate (if any) whose loss gives the most improvement of the fit
3. Repeating this process until no further covariates can be deleted without a loss of fit

(RMD_example 11.8)

Forward selection

1. Starting with no covariates in the model
2. Testing the addition of each covariate using a chosen model fit criterion, adding the covariate (if any) whose inclusion gives the most improvement of the fit
3. Repeating this process until none improves the model

(RMD_example 11.8)

Stepwise regression

- Start like forward selection
- A combination of forward and backward, testing at each step for covariates to be included or excluded.

(RMD_example 11.8)