ZZSC9001 Assessment 2 - Econometrics

z5428434 A. Gill

September 26, 2022

Introduction

Challenging economic conditions, influenced by uncertainty both locally and abroad (CoreLogic, 2022), have contributed to a recent decline in housing prices in Greater Melbourne (Real Estate Institute of Victoria, 2022). This multiple linear regression analysis aims to capture the effects of two variables - dwelling size and location - on the price of a family home in Greater Melbourne. For each property, these variables are fixed and thus immune to macroeconomic or global political changes, providing an opportunity for stable predictions during an otherwise turbulent period.

Question 1

A sample of 20 houses has been acquired for this study, via a **multi-stage** sampling strategy:

1. The housing stock in Greater Melbourne was **stratified** by its position relative to central Melbourne; that is, east and west. This was to control the possibility that this positioning might affect housing prices.

2. Within each of these strata, the housing stock was **clustered** by suburb. This required an assumption that every suburb (cluster) in the stratum was a fair representation of the characteristics of the entire stratum (Statistics How To, n.d.).

3. One suburb was selected from each stratum; Dandenong in the east, and Sunshine in the west. Both of these suburbs are characterised by considerable socioeconomic disadvantage (Australian Bureau of Statistics, 2016). This **non-random** selection of two suburbs with a similar socioeconomic profile should neutralise the effects of socioeconomic factors on the analysis.

4. Ten houses were **randomly** selected from each of the two suburbs. This required an assumption that a small sample is sufficient to represent each suburb and to observe a relationship between the variables.

Question 2

Before embarking on the regression analysis, some preliminary exploration was conducted in order to establish the suitability of the acquired data to linear regression. Note that the dwelling size has been approximated using the number of rooms.

Figure 1 shows a clear positive association between selling price and the number of rooms, confirming suitability to a linear regression analysis (Sharpe et al., 2014a). This is confirmed by a correlation coefficient of 0.81.

Figure 2a shows a well-defined separation between selling prices in Dandenong and Sunshine, indicating that location is relevant to selling price. Given that location is an indicator with only

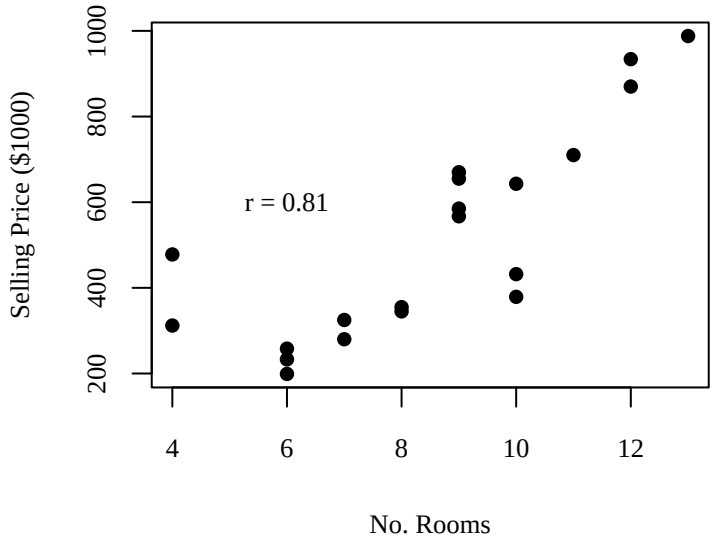
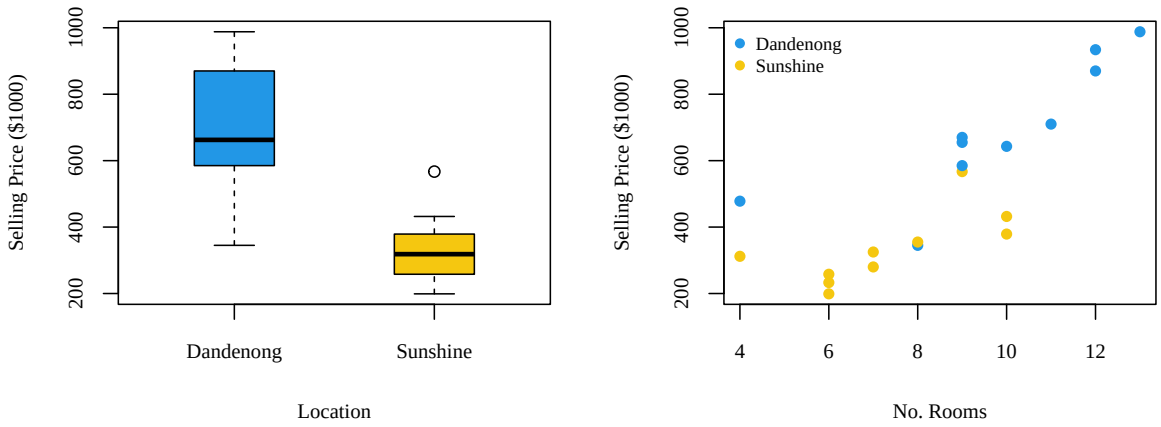


Figure 1: Selling Price vs No. Rooms

two values, a box plot was deemed appropriate for conveying the suitability of location to this regression analysis.

However, it appears from Figure 2b that the Dandenong sample is dominated by larger houses while the Sunshine sample is dominated by smaller houses, indicating a possible interaction between location and dwelling size (Sharpe et al., 2014b), or perhaps the influence of an unmeasured variable; this observation receives further attention in Question 6.



(a) Selling Price vs No. Rooms **(b) Selling Price by Location**

Figure 2: Effect of Location on Selling Price

Having established an association between selling price and the two predictors, the multiple linear regression was carried out, yielding the residuals shown in Figure 3. There appears to be fairly uniform variation across all predicted selling prices, and there is no apparent pattern, confirming that the model likely meets the linearity condition (Sharpe et al., 2014c).

Figure 4a shows the residuals plotted against the number of rooms. While there is some clustering along the middle band of the predictor, this is not to the extent that would warrant an adjustment to linearise the relationship (Sharpe et al., 2014d). However, Figure 4b shows that the model is a better predictor of house prices in Sunshine than in Dandenong; this is evident in the greater spread of residuals for Dandenong.

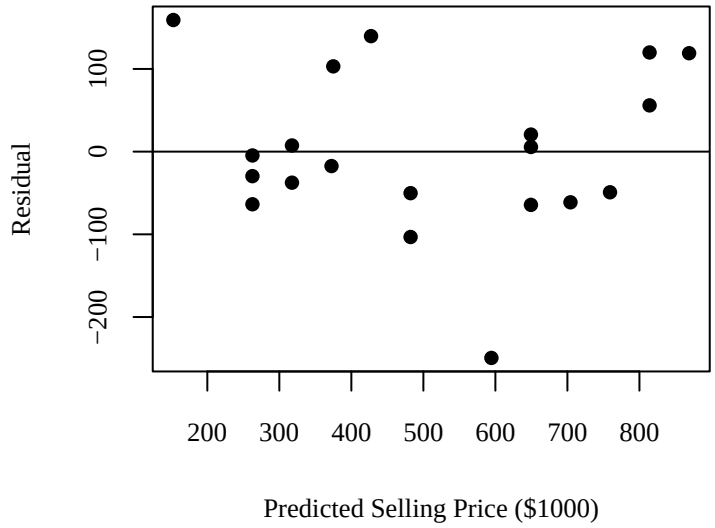
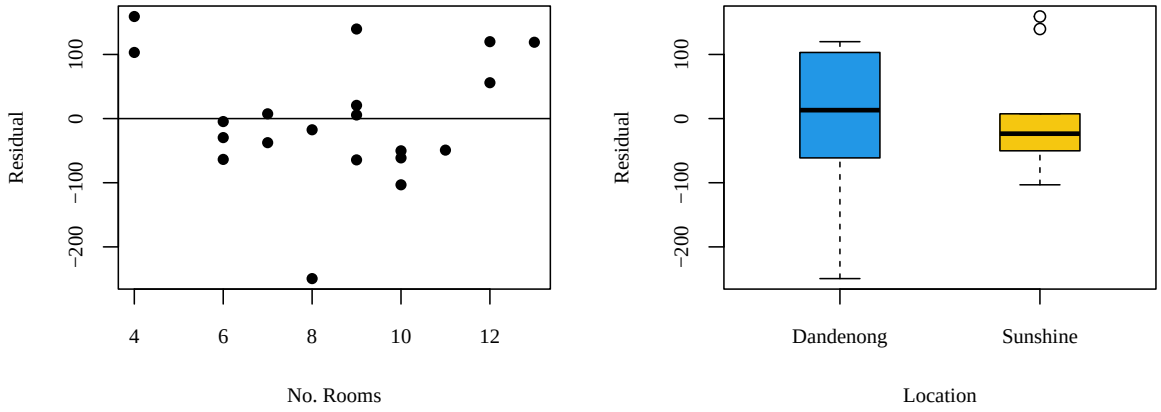


Figure 3: Residuals vs Predicted Selling Price



(a) Residuals vs No. Rooms **(b) Residuals by Location**

Figure 4: Residuals Against Predictor Variables

Question 3

The linear regression analysis was completed using R. The R script can be found in Appendix I, and its output can be found in Appendix II. The equation of the best-fit line is as follows.

$$price = 155.29 - 222.04 \times location + 54.90 \times rooms$$

The coefficients suggest that:

- When the location is 1 (i.e. Sunshine) and the number of rooms remains steady, the selling price **reduces** by $222,040.
- When the number of rooms increases by 1 and the location remains steady, the selling price **increases** by $54,900.

This is consistent with the observations made in Question 2, on Figures 1 and 2a.

If a prediction were to be made of the selling price for a nine-room house in Melbourne's east (i.e. Dandenong), the regression equation would yield the following result.

$$price = 155.29 - 222.04 \times 0 + 54.90 \times 9 = \$649,390$$

Question 4

The reported F-statistic is 42.95 with a p-value of 2.26×10^{-7}.

To exceed a 5% level of significance, the p-value must be less than 0.05 (Pennsylvania State University - Department of Statistics, n.d.). The results show that this is indeed the case. This suggests that the null hypothesis, which states that the coefficents of the predictor variables are 0 and that this model is likely no better at predicting the selling price than the mean (Sharpe et al., 2014e), can be **rejected**.

This confirms a jointly significant relationship between the selling price and the two predictors, location and number of rooms.

Question 5

Examination of the predictors' individual contributions to the predicted selling price yields the following observations:

- For the location variable, the t-statistic is -4.222 and the p-value is 5.74×10^{-4}. This is below the 0.05 benchmark for 5% significance; therefore, it can be said that location is a significant predictor of selling price.
- For the number of rooms, the t-statistic is 5.177 and the p-value is 7.58×10^{-5}. This is much smaller than the 0.05 required for 5% significance; therefore, it can be said that the number of rooms is a significant predictor of selling price.

Question 6

The regression analysis has been repeated, this time with an interaction term, $location \times rooms$. This adjustment to the model should validate (or otherwise) the potential interaction that was observed in Figure 2b. The updated regression equation takes the following form.

$$price = 61.997 - 5.011 \times location + 64.516 \times rooms - 26.569 \times location \times rooms$$

The t-statistic for the interaction term is -1.221, and the p-value is 0.240. These results suggest that the interaction between location and number of rooms is **not** a significant predictor of selling price. Therefore, some other explanation likely exists for the separation observed in Figure 2b.

Question 7

Given the results discussed in Question 6, it would be natural to conclude that the **original** multiple regression is the preferable model. However, there are observations within the interaction model that are worth examining before a final determination is made.

The *adjusted* coefficient of determination, R^2, for the original model is 0.815. For the interaction model it is 0.821. This suggests that for both models, approximately 82% of the variability in selling price can be explained by the variability in their respective predictor variables (Staggard, 2015).

However, the F-statistic for the interaction model is 29.96. This is lower than that for the original model. While the p-value of the interaction model is much lower than required for 5%

significance, it is still higher than that of the original model, and its lower F-statistic compares unfavourably with the original.

Therefore, even after closer examination, the original regression model remains the preferred.

Conclusion

Small samples of housing prices in eastern and western Melbourne have been used in a multiple linear regression analysis, establishing that the selling price of a house can reasonably be predicted from its size (measured as number of rooms) and location.

Additional modelling has also been completed to rule out the significance of any interaction between location and number of rooms. The results of this modelling confirm that the original model should stand.

References

Australian Bureau of Statistics (2016). *Socio-Economic Indexes for Areas (SEIFA)*. https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/2033.0.55.001Main+Features12016?OpenDocument. [Accessed on 22-09-2022].

CoreLogic (2022). *What does high inflation mean for the Australian housing market?* https://www.corelogic.com.au/news-research/news/2022/what-does-high-inflation-mean-for-the-australian-housing-market. [Accessed on 22-09-2022].

Pennsylvania State University - Department of Statistics (n.d.). *STAT415: Introduction to Mathematical Statistics - Lesson 10.2*. https://online.stat.psu.edu/stat415/lesson/10/10.2. [Accessed on 21-09-2022].

Real Estate Institute of Victoria (2022). *Market Insights - Metropolitan Melbourne*. https://reiv.com.au/market-insights/victorian-insights. [Accessed on 22-09-2022].

Sharpe, N.D. et al. (2014a). *Business Statistics*. Pearson Education. Chap. 4.1.

— (2014b). *Business Statistics*. Pearson Education. Chap. 18.2.

— (2014c). *Business Statistics*. Pearson Education. Chap. 15.2.

— (2014d). *Business Statistics*. Pearson Education. Chap. 16.6.

— (2014e). *Business Statistics*. Pearson Education. Chap. 15.1.

Staggard, K. (2015). *CSM VCE Further Mathematics Units 3 and 4*. Cambridge University Press. Chap. 4.

Statistics How To (n.d.). *Cluster Sampling in Statistics: Definition, Types*. https://www.statisticshowto.com/what-is-cluster-sampling/. [Accessed on 22-09-2022].

Appendix I: R Code

```
# ZZSC9001 ASSESSMENT 2: ECONOMETRICS
# AUTHOR: z5428434 A. Gill

# install package for reading xlsx files
# uncomment if not already installed
# install.packages('readxl')

# load library
library(readxl)

# read xlsx data
houses <- read_xlsx('provided info/Housing prices.xlsx')
houses

# rename variables for readability
price <- houses$`Selling Price`
rooms <- houses$`Number of Rooms`
location <- houses$Location

# globally set display features for all graphs
par(
    family='serif',
    ps=10.5,
    mar=c(4.5, 4.5, 1, 1)
)

# plot 1 - selling price vs no. rooms
plot(
    rooms, price,
    pch = 16,
    xlab = 'No. Rooms',
    ylab = 'Selling Price ($1000)',
)

cor.price.rooms <- round(cor(price, rooms), 2)

text(
    x = 6,
    y = 600,
    label = paste('r =', cor.price.rooms),
    vfont=NULL
)

# plot 2 - boxplot for selling price by location
boxplot(
    price ~ location,
    col = c(4, 7),
    xlab = 'Location',
    ylab = 'Selling Price ($1000)',
```

```
    names = c('Dandenong', 'Sunshine'),
    boxwex = 0.4
)

# plot 3 - selling prices vs no. rooms by location
plot(
    rooms, price,
    col = (location*3)+4,
    pch = 16,
    xlab = 'No. Rooms',
    ylab = 'Selling Price ($1000)',
)

legend(
    x = 'topleft',
    legend = c('Dandenong', 'Sunshine'),
    col = c(4, 7),
    pch = 16,
    bty = 'n',
    cex = 0.9
)

# multiple linear regression with 2 explanatory variables
reg_model <- lm(price ~ location + rooms)
summary(reg_model)

# residuals and predicted selling prices
resids <- resid(reg_model)
preds <- fitted(reg_model)

# plot 4 - residuals vs predicted selling price
plot(
    preds, resids,
    pch = 16,
    xlab = 'Predicted Selling Price ($1000)',
    ylab = 'Residual',
)

abline(0, 0)

# plot 5 - residuals vs no. rooms
plot(
    rooms, resids,
    pch = 16,
    xlab = 'No. Rooms',
    ylab = 'Residual',
)

abline(0, 0)

# plot 6 - residuals by location
```

```
boxplot(
    resids ~ location,
    col = c(4, 7),
    xlab = 'Location',
    ylab = 'Residual',
    names = c('Dandenong', 'Sunshine'),
    boxwex = 0.4
)

# prediction for 9-room house in Dandenong
pred.9rms.loc0 <- predict(reg_model, list(location=0, rooms=9))
pred.9rms.loc0

# multiple linear regression with an interaction term
reg_model.interaction <- lm(price ~ location + rooms + location:rooms)
summary(reg_model.interaction)
```

Appendix II: Regression Output

Original model with two explanatory variables:

```
Call:
lm(formula = price ~ location + rooms)

Residuals:
    Min      1Q  Median      3Q     Max
-249.47  -52.99  -11.03   67.73  159.16

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   155.29     107.88   1.439 0.168191
location     -222.04      52.59  -4.222 0.000574 ***
rooms          54.90      10.60   5.177 7.58e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 102.9 on 17 degrees of freedom
Multiple R-squared:  0.8348,    Adjusted R-squared:  0.8153
F-statistic: 42.95 on 2 and 17 DF,  p-value: 2.257e-07
```

Adjusted model with an interacton term:

```
Call:
lm(formula = price ~ location + rooms + location:rooms)

Residuals:
    Min      1Q  Median      3Q     Max
-233.12  -57.50   -5.01   47.19  168.49

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      61.997    130.940   0.473 0.642271
location         -5.011    185.099  -0.027 0.978736
rooms            64.516     13.087   4.930 0.000151 ***
location:rooms  -26.569     21.752  -1.221 0.239620
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101.5 on 16 degrees of freedom
Multiple R-squared:  0.8489,    Adjusted R-squared:  0.8205
F-statistic: 29.96 on 3 and 16 DF,  p-value: 8.452e-07
```