ECON 453, Spring 2023, Pset 3
Minh Duong

Problem Set 3 Report

Minh Duong

04 – 09 – 2023

a. Estimated Equations and Meaning of estimated parameters:

$$Price = 129397.10 + 66.72 * Area - 1138.64 * Age + 5859.00 * Pool + 9287.42 * Northern$$

The interpretation of the model parameters is as follows:

- The intercept parameter (129397.10) represents the predicted price of a house when all other input variables are equal to zero. In other words, it represents the base price of a house without considering any of the other factors.
- The parameter for the Area variable (66.72) represents the change in price that is associated with a one-unit increase in the Area variable, while holding all other variables constant. This means that for each additional unit of area, the predicted price of the house increases by 66.72 units, assuming all other variables remain constant.
- The parameter for the Age variable (-1138.64) represents the change in price that is associated with a one-unit increase in the Age variable, while holding all other variables constant. This means that for each additional year of age, the predicted price of the house decreases by 1138.64 units, assuming all other variables remain constant. The age is calculated in the year **2023**.
- The parameter for the Pool variable (5859.00) represents the change in price that is associated with having a pool, compared to not having a pool, while holding all other variables constant. This means that, on average, having a pool increases the predicted price of the house by 5859.00 units, assuming all other variables remain constant.
- The parameter for the Northern variable (9287.42) represents the change in price that is associated with being in the Northern region, compared to being in the Southern region, while holding all other variables constant. This means that, on average, being in the Northern region increases the predicted price of the house by 9287.42 units, assuming all other variables remain constant.

Overall, the model suggests that the most significant predictors of house price are Area, Age, Pool, and Northern region. The coefficients for these variables indicate the direction and strength of their impact on the predicted price of the house.

b. Does the location of the house matter, i.e., is there a premium for a house located in the prestigious "northern" part of the town?

The model gives the parameter for the Location variable of 9287.4, with a p-value of 0.0557.

The p-value for the parameter is 0.0557, which is larger than the typical significance level of 0.05. This means that there is a 5.57% chance that the observed effect of the Northern variable on the price of the house is due to random chance, assuming that there is actually no effect of the Northern variable on the price.

Since the p-value is slightly higher than the typical significance level of 0.05, we cannot conclusively say that there is a significant effect of the Northern variable on the price of the house. However, the p-value is close to 0.05, which suggests that there may be some evidence of an effect.

c. Does the presence of a swimming pool influence the selling price of the house?

The parameter for the Pool variable has a p-value of 0.2384, which is larger than the typical significance level of 0.05. This means that there is a 23.84% chance that the observed effect of the Pool variable on the price of the house is due to random chance, assuming that there is actually no effect of the Pool variable on the price. Since the p-value is higher than the typical significance level of 0.05, we cannot conclude that there is a significant effect of the Pool variable on the price of the house.

d. Does your estimated equation support your hypothesis that "size of house matters?" Interpret the parameter associated with "house size" variable.

The model gives the p-value for the parameter associated with the "Area" (house size variable) of $1.18e-8$. Such p-value is less than the commonly used significance level of 0.05. This indicates that the "Area" variable is statistically significant in predicting the price of the house.

The parameter associated with the "Area" variable is 66.72, which represents the estimated change in the price of the house for every one-unit increase in the "Area" variable, while holding all other variables constant. In other words, this means that for every additional square unit increase in the area of the house, the price of the house is estimated to increase by $66.72, holding all other factors constant.

e. Mr. Dimeworth believes that newer houses fetch a greater price, other things being equal. Test his hypothesis.

The p-value for the "Age" parameter is $1.69e-6$, which is less than the commonly used significance level of 0.05. This indicates that the "Age" variable is also statistically significant in predicting the price of the house. The parameter estimate for "Age" in the given model is -1138.64, which means that holding all other variables constant, the price of the house is estimated to decrease by $1138.64 for every one-year increase in the age of the house.

An additional calculation is made to test the hypothesis of Mr. Dimeworth, with H_0 being the parameter for Age is larger than or equal to 0, and H_1 being the parameter for Age is smaller than 0. The test results in the p-value of $8.43e-7$, which is less than 0.05. The result suggests that the age of the house has a statistically significant effect on lowering the price of the house. This also means that the newer the house, the lower the price. Mr. Dimeworth's hypothesis is thus justified.

f. Do a test of the overall significance of the regression equation. State your null hypothesis and conclusion clearly.

To test the overall significance of the regression equation, we have the the hypothesis:

- H_0: all of the parameters of the independent variables are 0 (param(Age) = 0; param(Area) = 0; param(Pool) = 0; param(Location) = 0)
- H_1: not H_0, which means at least 1 of the parameters is not 0.

We have 4 independent variables and 50 observations, so we will compare F_{calc} with $F_{4, 45}$. The result using R gives the test statistic (F value) to be 25.66, and the p-value of $4.086e-11$, which is lower than 0.05. That means we reject the null hypothesis, and the regression equation is statistically significant.

g. What is your best forecast of the price for a 2,000 square foot house built in 1990 with a swimming pool, a fireplace, and a garage?

We have the model:

$$Price = 129397.10 + 66.72 * Area - 1138.64 * Age + 5859.00 * Pool + 9287.42 * Northern$$

We can also use the predict() function in R, which is the way to get the result from the model.

- For the Northern location, the price is (by R): 240398.4
- For the non-Northern location, the price is (by R): 231111

h. Considering everything, explain why you think you have a (un)reasonable model?

Based on the result from R, we can analyze the model's overall statistical significance, goodness-of-fit, and the individual significance of the predictor variables.

Starting with the p-value for the whole equation, which is 4.086e-11. This suggests that the model is statistically significant since the p-value is much smaller than the commonly used significance level of 0.05. This indicates that at least one of the predictor variables is significantly related to the outcome variable (i.e., house price) in the model.

Moving on to the adjusted R-squared, which is 0.6681, this indicates that the model explains about 66.81% of the variation in the outcome variable. This value is not too high nor too low, but it suggests that the model may have some room for improvement in explaining the variation in the outcome variable. However, it is important to consider the context of the problem and what the R-squared value means for the specific application.

Regarding the individual predictor variables, we can look at their respective coefficients and p-values to determine their significance. The Area and Age variables have parameters with p-value of less than 0.05, which indicates that they are each statistically significant in predicting the outcome variable, given that the other variables are held constant. The Pool and Location (dummy) variables does not have a p-value of less than 0.05, however, which suggests that they are not statistically significant in determining the house price.

In summary, based on the provided information, the model appears to be reasonably good at predicting house prices, since it is statistically significant as a whole, has an adjusted R-squared of 0.6681, and the coefficients for some predictor variables are statistically significant.

i. You present your results to Mr. Dimeworth. Trying not to let his admiration for your work out, he asks, "What does your model tell us about the value of a garage?" Can you use your estimated equation to impute the value of a garage? If not, how do you modify your model so you can measure the value of a garage?

The estimated equation currently only includes four predictor variables: Area, Age, Pool, and Northern. Therefore, the model does not directly provide information on the value of a garage. However, we can modify the model to include a new predictor variable for the presence of a garage and estimate its coefficient to measure the value of a garage in the model. To do this, we would need to obtain data on the presence of a garage for the same 50 observations used to estimate the current model. We could then add a new dummy variable to the model, "Garage," and assign a binary value of 0 or 1 depending on whether each observation has a garage or not. We could then estimate the new coefficient for "Garage" to determine the effect of having a garage on the house price, while holding the other variables constant.

The modified model would look like:

$$Price = \beta 0 + \beta 1 * Area + \beta 2 * Age + \beta 3 * Pool + \beta 4 * Northern + \beta 5 * Garage + \varepsilon$$

where β5 is the coefficient for the new variable "Garage." Once the model is estimated with the new variable included, we could use the coefficient estimate for "Garage" to impute the value of a garage on the house price.

As a side note here, it is important to note that adding a new predictor variable may affect the coefficients and statistical significance of the other variables in the model. Therefore, it is important to re-evaluate the modified model's goodness-of-fit, statistical significance, and other assumptions before drawing any conclusions about the value of a garage based on the model.

j. How do you modify your equation so that the effect of house size on price depends on the location? For full credit, estimate a model that allows the effect of house size on house price to vary across the two locations. Write your estimated equation(s) and interpret key parameter estimates. Test the hypothesis that the effect of house size on price is the same for both locations (North and not-North).

To include the effect of house size on price depending on the location, we can add the term $Area * Northern$ into our model, and re-run the regression. In R, to do so, we can create a new variable Area_Location, which is the multiplication of Area value and Location (Northern dummy variable) values, and include it in our new regression model. The resulting equation is:

$$Price = 140834.10 + 62.01 * Area - 1184.54 * Age + 6579.52 * Pool - 28688.40 * Northern + 19.33 * Area * Northern$$

This model can be interpreted as:

- The intercept of the model is 140,834.10. This represents the estimated house price when all other independent variables in the model are set to zero.
- The coefficient of the "Area" variable is 62.01, indicating that for every additional square foot of area, the house price is estimated to increase by $62.01, holding all other independent variables constant.
- The coefficient of the "Age" variable is -1184.54, indicating that for every additional year of age of the house, the house price is estimated to decrease by $1184.54, holding all other independent variables constant.
- The coefficient of the "Pool" variable is 6579.52, indicating that if the house has a pool, the house price is estimated to increase by $6579.52, holding all other independent variables constant.
- The coefficient of the "Northern" variable is -28688.40, indicating that if the house is located in the northern region, the house price is estimated to decrease by $28,688.40, holding all other independent variables constant.
- The coefficient of the "Area*Northern" interaction variable is 19.33, indicating that the effect of "Area" on the house price is modified by the "Northern" region. Specifically, for each additional square foot of area, the house price is estimated to increase by $19.33 more in the northern region than in the southern region, holding all other independent variables constant.

To test the hypothesis that the effect of house size on price is the same for both locations, we have the null hypothesis that the effect of house size on price is the same (parameter of Area * Northern = 0), and the alternative hypothesis is it is not equal to 0. The calculation from R gives us a p-value of 0.414, which is higher than the 0.05 significance level. It means that the effect of house size on the price of the two regions are not significantly different, or in other words, they are the same.

k. To do so in R, we can create another variable, which is equal to the Price variable divided by 1000, and use that variable as the dependent variable in our model instead.

The result indicates that the parameters (both the slope and intercept parameters) are also divided by 1000 from the original model, with their standard errors also being divided by 1000. This can be explained from the change in unit, as the unit of the dependent variable is multiplied by 1000 (dollar to thousands of dollars), its value is divided by 1000. The units of the parameters for each independent

variable thus are multiplied by 1000, which resulted in the values being divided by 1000. As the standard error has the same unit as the mean, it is divided by 1000 as well.

However, the test statistics (t-value, f-value), p-value and R-squared do not change. This is because these values do not have a unit. As those statistics are standardized to be unit-less and they represent the statistical significance of the independent variables' parameters, their explanatory proportions as well as the overall reasonability of the model, the change in the unit of the dependent variable does not create any impact on them, and thus they are unchanged.

Therefore, while the estimated coefficients are divided by 1000, the significance of the parameters and R-squared does not change.

ECON 453, Spring 2023, Pset 3
Minh Duong

Addendum

Project Repository: https://github.com/MykeDuong/econ453

R Script code:

```
## ----setup, include=FALSE-----------------------------------------
knitr::opts_chunk$set(echo = TRUE)


## ------------------------------------------------------------------
rm(list = ls()) # clear the workspace

getwd()
setwd("./")

library(readxl)
library(car) # loads car library that does tests of linear restrictions

data <- read_excel("./data/pset3_data.xlsx")

# only get data from row 2-51 (-1 for 0-indexed)
data = data[c(1:50), ]
# Convert to numerical data typee
data$Price <- as.numeric(data$Price)
data$Area <- as.numeric(data$Area)
# Convert dummy variables
data$D_Pool = ifelse((data$Pool == "yes"), 1, 0)
data$D_Location = ifelse((data$Location == "North"), 1, 0)
# Convert year (19xx) to age
data$Age= 2023 - (1900 + data$Year) # Change to 2023

summary(data)

# 1A. Estimated equation
model <- lm(data = data, Price ~ Area + Age + D_Pool + D_Location)
summary(model, digits = 5)

# 1B. House location
model$coefficients["D_Location"]

# 1C. Swimming pool
model$coefficients["D_Pool"]

# 1D. House Size
model$coefficients["Area"]

# 1E. House Age
model$coefficients["Age"]
## 1-tailed test
pt(coef(summary(model))[, 3], model$df, lower = TRUE)

# 1F. Overall significance
coefs <- names(coef(model))
```

```
linearHypothesis(model, coefs[-1])

linearHypothesis(model, c(
  "Age = 0",
  "Area = 0",
  "D_Pool = 0",
  "D_Location = 0"
))

# 1G. Forecast of the price for a 2,000 square foot house built in 1990 with
#     a swimming pool, a fireplace, and a garage?
forecastNorthernValue <- data.frame(
  Age = c(2023 - 1990),
  Area = c(2000),
  D_Pool = c(1),
  D_Location = c(1)
)

predict(model, forecastNorthernValue)

forecastNonNorthernValue <- data.frame(
  Area = c(2000),
  Age = c(2023 - 1990),
  D_Pool = c(1),
  D_Location = c(0)
)

predict(model, forecastNonNorthernValue)

# 1H. Reasonable

# 1I.

# 1J.
data$Area_Location = data$Area * data$D_Location

priceAreaModel = lm(
  data = data,
  Price ~ Area + Age + D_Pool + D_Location + Area_Location
)

summary(priceAreaModel)

# 1K.
data$PriceK = data$Price / 1000
kModel = lm(data = data, PriceK ~ Area + Age + D_Pool + D_Location)
summary(kModel)
```