ECON 213A – Applied Econometrics with R
Thursday, April 15th
Professor Ben Koskinen

Answers to Mid-term Questions

Zilin Luo

“I, Zilin Luo, understand that this midterm is to be completed independently. I have asked no questions to any other human, with the exception of the TA or the professor, regarding this midterm. I understand that any suspicion of plagiarism or academic dishonesty will result in a zero grade for this midterm. This Statement and my name act as an understanding and compliance of the terms. Signed, Zilin Luo.”

1. Data Inspection and Statistics Inference

(a)

This dataset contains 27 variables and 25357 observations.

Garage, sold93, sold94, sold95, sold96, sold97, sold98 are dummy variables.

Wall and garage_type are categorical variables.

(b)

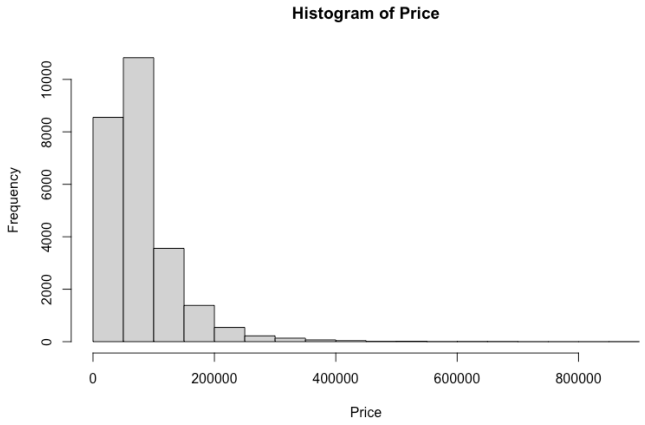

Figure 1 Histogram of Price

The data in the graph are right-skewed. This indicates that the sale price of the most house is less than 200000, especially concentrated in the price between 0 and 10000. Houses with sale prices ranging from 5,000 to 10,000 are the most (over 10000), and the number of homes decreases with the rise of the sale price.

(c)

The correlation between the square footage of the living area and price is 0.7682, showing that the larger the living area is, the higher the sale price is.

Price is affected by different types of garage. Figure 2 shows that each type of garage has a different range and average price. Generally, the price of a house with an attached garage is higher than houses with other types of garage. Additionally, the p-value of ANOVA also shows that the average price would change in terms of different types of garage.

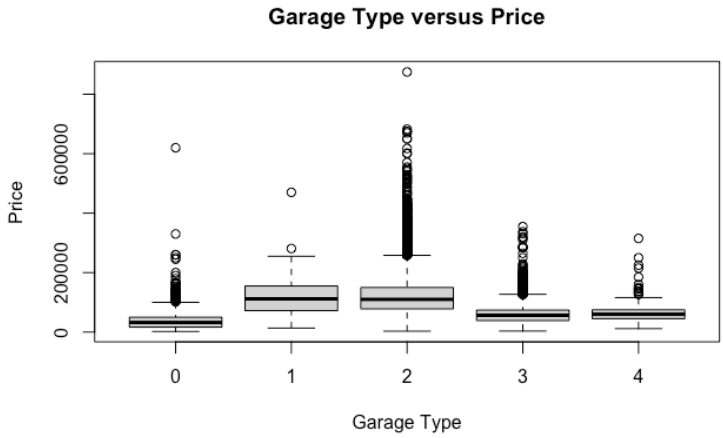

Figure 2 Garage Type versus Price

(d)

In the histogram of the sale date, bins are divided into months. Compared with the year as histogram interval, month classes display the number of house transactions not only in the year level but also in the month level. And also, grouping sale dates by days or weeks would cause much more bars in the histogram, making people difficult to obtain overall trends during the whole period.

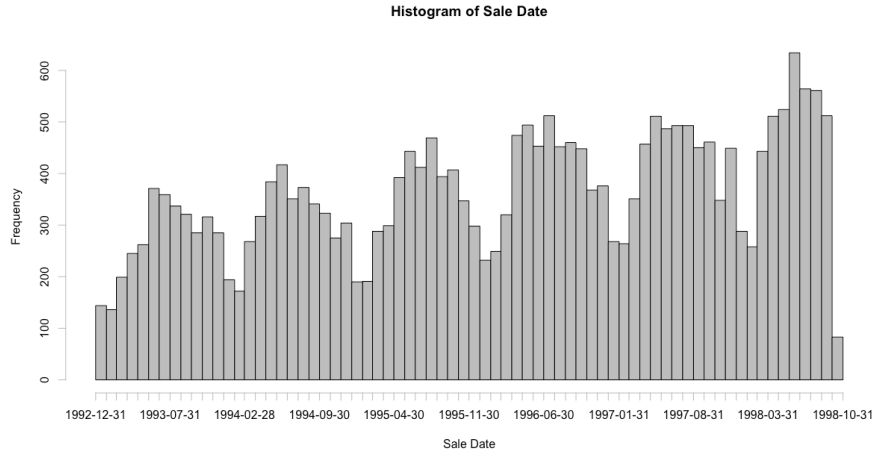

Figure 3 Histogram of Sale Date

Figure 3 shows that the trend of sales is similar in each year. Overall, there is an upward trend from the first quarter to the second quarter, and then a downward trend in the following months. The biggest number of houses are sold in summer, fewer people buy a house in winter.

2. Model Selection and Output

(a)

$$\text{Price} = \beta_0 + \beta_1\text{yrbuilt} + \beta_2\text{stories} + \beta_3\text{sqft} + \beta_4\text{sqft}^2 + \beta_5\text{wall} + \beta_6\text{bed} + \beta_7\text{bath} + \beta_8\text{bath}^2 + \beta_9\text{frontage} + \beta_{10}\text{garage\_type} + \beta_{11}\text{garage\_sqft} + \beta_{12}\text{distance} + \beta_{13}\text{saleq}$$

In this model, the price of houses are affected by 13 variables, including the year the house was built, number of stories, area of a living area (square footage), the square of the area of living area, the construction material of the wall, number of bedrooms, number of bathrooms, the square of the number of bathrooms, frontage of the house, the type of garage, area of the garage (square footage), physical distance to the city center, the quarter the house sold.

These are the independent variables that were chosen:

Independent Variable Description			
Variable Name	Prior	Variable Name	Prior
Year Build	$\beta 1 > 0$	Number of Bathrooms^2	$\beta 8 > 0$
Stories	$\beta 2 > 0$	Frontage of House	$\beta 9 > 0$
Area of Living Area	$\beta 3 > 0$	Type of Garage	$\beta 10 < 0$
Area of Living Area^2	$\beta 4 > 0$	Area of Garage	$\beta 11 > 0$
Wall Material	$\beta 5 > 0$	Distance from House to City Center	$\beta 12 < 0$
Number of Bedrooms	$\beta 6 > 0$	The Quarter the House Sold	$\beta 13 > 0$
Number of Bathrooms	$\beta 7 > 0$		

Table 1 Independent Variable Description

There is no variable added from external sources. Four variables are created as a transformation of other variables. They are:

1. The square of the area of a living area ($sqft^2$): In the scatter plot of area of living area and the difference between estimated price and real sale price, when the area increases, the difference between the estimated value and the true value is not constant but continuously expanding. Some omitted or unobservable variables likely have an impact on the sale price of a house.
2. The square of the number of bathrooms ($bath^2$): As there is a problem on non-constant variance on error term, and those are the variable seems to related to the non-constant variance as they cannot estimate the dependent variable efficiently, so a square term is added to here to see if it would help improve the fit of the model.
3. Distance from House to City Center (distance): Given that the area close to the city center will be more prosperous than areas away from the city center in terms of population density (more demand for a living) and easy access to different resources like shopping malls, office buildings, hospitals and schools, the selling price of houses per squared area is proposed to higher than suburbs. In order words, the distance between houses and the city center has a significant impact on the price of a house.

   Toledo is the city center of Lucas County. The latitude and longitude of Toledo are (41.6528, -83.5379), based on the latitude and longitude of Toledo and houses, the difference in latitude and longitude could be obtained. With the conversion relationship between miles and latitude and longitude provided, the distance from house to city center could be calculated by using the distance formula.
4. The Quarter the House Sold (sales): The histogram of the date of house sold denotes different quarter have an effect on the sale price. It seems that more people buy houses in summer than in other seasons. The number of house transactions is related to the quarter, so a new variable is created based on the date of house sold.

Reasons for choosing variables and expected effect of each variable on sale price:

1. Year Build: In general, under the same conditions, the new house is more valuable than the old one. It is expected that the sale price of houses would increase with the year built.
2. Stories: More stories constructed means higher construction costs spent, resulting in higher prices. On the other hand, the price is also decided by supply and demand. House stories with 2 or 3 stories are likely more popular than houses with other numbers of floors, leading to the

different sale prices for different types of stories. It is expected that the price will first increase as the number of stories increases and then go down with the more stories.

3. Living Area: The living area excludes outside areas of the house such as gardens. The area that a lot has designed for the living building means carries significant construction costs like labor and material. It is expected that a larger living area will have a larger selling price.
4. Wall Material: Different types of material like wood, steel bring involve different construction costs, having different impacts on price. It is expected that more expensive material used would cause a higher sale price of that house.
5. Number of Bedrooms: To some extent, houses with more bedrooms cost more. It is expected that the number of bedrooms has a positive effect on the price.
6. Number of Bathrooms: Same as the number of bedrooms, it is likely that the number of bathrooms has a positive impact on the sale price.
7. Frontage of House:
8. Type of Garage: This variable not only involves information that whether the house has a garage but also provides information on the type of garage. It is expected that some kinds of the garage are more welcomed by individuals, resulting in higher house prices.
9. Area of Garage: It is estimated that houses with larger garages are more costly than smaller garages.
10. Distance from house to City Center: Different areas have different prices for land because of demographic factors. It would be naturally expected that the distance between a house and the city center have a negative impact on price.
11. The Quarter the House Sold: the price is significantly decided by the demand and supply side. In winter, fewer people tend to buy houses, the number of houses on the market is greater than needed, house prices would likely drop in that season.

(b)

This dataset does not include the population of Lucas County, population growth, income, interest rate, construction costs, supply of other kinds of houses. These variables may change the supply and demand of housing, affecting the price of houses.

All variables above may cause omitted variable bias.

(c)

Heteroskedasticity is present. According to the residual and variable plot.

The image of residua and variables shows that house price fluctuates greatly under the influence of variables.

(d)

Model Selection:

1. Model 1:

$$\text{Price} = \beta_0 + \beta_1\text{yrbuilt} + \beta_2\text{stories} + \beta_3\text{sqft} + \beta_4\text{wall} + \beta_5\text{bed} + \beta_6\text{bath} + \beta_7\text{frontage} + \beta_8\text{garage\_type} + \beta_9\text{garage\_sqft} + \beta_{10}\text{lotsize} + \beta_{11}\text{distance} + \beta_{12}\text{saleq}$$

Firstly, I chose the variables that I thought have an impact on housing price.

2. Model 1.1

$$\text{Price} = \beta_0 + \beta_1\text{yrbuilt} + \beta_2\text{stories} + \beta_3\text{sqft} + \beta_4\text{wall} + \beta_5\text{rooms} + \beta_6\text{frontage} + \beta_7\text{garage\_type} + \beta_8\text{garage\_sqft} + \beta_9\text{lotsize} + \beta_{10}\text{distance} + \beta_{11}\text{saleq}$$

Given that the number of rooms is highly correlated with the number of bedrooms and the number of bathrooms. The number of bedrooms and bathrooms are included in the number of rooms. To

avoid collinearity between variables, In Model 1.1, variables the number of bedrooms and the number of rooms are replaced by the number of rooms. Compared with Model 1.1, Model 1 has a higher R^2 (0.730), suggesting that the price estimated by model 1 is closer to its true value. So based on model 1 we do further analysis.

3. Model 1.2

$$Price = \beta_0 + \beta_1 yrbuilt + \beta_2 stories + \beta_3 sqft + \beta_4 wall + \beta_5 bed + \beta_6 bath + \beta_7 frontage + \beta_8 garage + \beta_9 garage\_sqft + \beta_{10} lotsize + \beta_{11} distance + \beta_{12} saleq$$

There are three variables related to garage in the original dataset, variable the type of garage contains the information variable whether the house has a garage, so these two variables could not be included in the model at the same time. In model 1.2, variable type of garage is replaced by the variable whether the house has a garage, model 1 and model 1.2 have very close R^2 (around 0.730), I still choose model 1 because the type of garage could convey more information than whether the house has a garage.

All variables needed are selected, and all values of the VIF of model 1are less than 5 demonstrate that there is no high collinearity among independent variables, which means that the change of anyone independent variable would not cause other independent variables to change.

4. Model 2

A very low p-value of the reset test indicates that at least the square form of one independent variable needs to be added to the model. By checking the scatter plot of variables and the difference between estimated price and teal price, variables the area of garages and the number of bathrooms were selected because there are clear patterns shown in their plots(Figure 4). The more bathrooms the house have, the

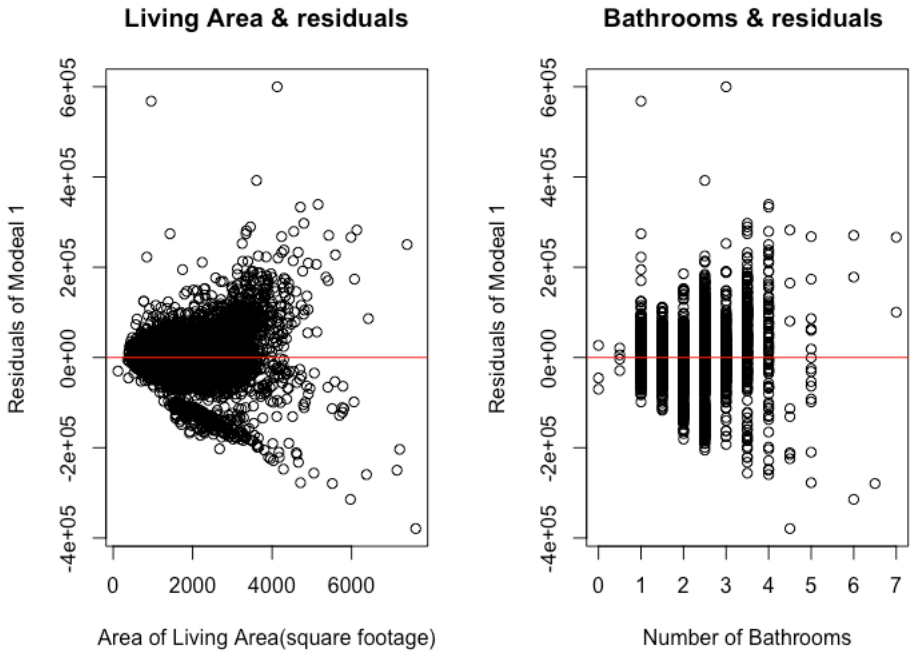

Figure 4

Then we change our model to model 2:

$$Price = \beta_0 + \beta_1 yrbuilt + \beta_2 stories + \beta_3 sqft + \beta_4 sqft^2 + \beta_5 wall + \beta_6 bed + \beta_7 bath + \beta_8 bath^2 +$$

$\beta_9$frontage + β_{10}garage_type + β_{11}garage_sqft + β_{12}lotsize + β_{13}distance + β_{14}saleq

This model shows a better R^2 of 0.738, showing estimated house price is closer to real price.

5. Model 2.1

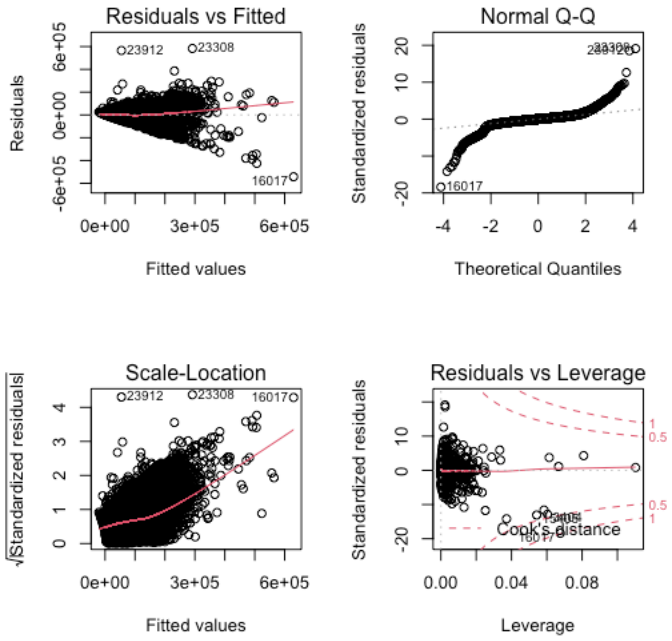

Figure 5 Diagnostic Plots

The image in the upper left corner shows that house price fluctuates greatly under the influence of variables. An image named Residuals vs Leverage in the bottom right corner indicates an influential outlier of 16017, which is needed to remove in mode 2.1.

the p-value of the variable lot size shows that this variable is 0.2039, which means that area of lot size has no impact on house price. I assumed that when the lot size of a house increases by one square foot, the house price would rise by 20.39%. A p-value of 0.5746 rejects that hypothesis. In other words, we could remove the variable area of lot size, then model 2.1 was created.

6. Model 2.2

A p-value of variable lot size shows that this variable is 0.2039, which means that area of lot size has no impact on house price. I assumed that when the lot size of a house increases by one square foot, the house price would rise by 20.39%. A P-value of 0.5746 rejects that hypothesis. In other words, we could remove the variable area of lot size, then model 2.2 was created. The R^2 of model 2.2 is 0.7415, denoting that the house price estimated is closer to real value compared to all previous models. I choose model 2.2 as the final model.

Price = β_0 + $\beta_1$yrbuilt + $\beta_2$stories + $\beta_3$sqft + $\beta_4$sqft^2 + $\beta_5$wall + $\beta_6$bed + $\beta_7$bath + $\beta_8$bath^2 +
$\beta_9$frontage + β_{10}garage_type + β_{11}garage_sqft + β_{12}distance + β_{13}saleq

Different regression results can be seen in table 2:

	Dependent variable:			
	House Price			
	Model 1	Model 2	Model 2.1	Model 2.2
Year Built	458.6349***	466.8026***	467.0385***	466.4089***
	(10.8325)	(10.7059)	(10.6344)	(10.6222)

Stories	929.6871***	1,758.1300***	1,851.0490***	1,854.3610***
	(133.0795)	(135.1980)	(134.3879)	(134.3624)
Living Area(square footage)	48.9161***	21.5212***	16.3440***	16.3281***
	(0.6267)	(1.4681)	(1.4848)	(1.4848)
Living Area(square footage)2		0.0056***	0.0069***	0.0069***
		(0.0003)	(0.0003)	(0.0003)
Wall2:Concrete block or tile	-11,436.2200**	-10,837.3800**	-10,539.9300**	-10,531.0700**
	(3,496.0960)	(3,446.9480)	(3,423.9450)	(3,423.9720)
Wall3:Aluminum,vinyl,or steel siding	-431.9063	609.9087	967.3615	972.2860
	(2,229.8620)	(2,198.7940)	(2,184.1820)	(2,184.2000)
Wall4:Brick	6,818.7790**	9,960.0520***	10,590.5500***	10,573.2800***
	(2,244.1910)	(2,215.6390)	(2,201.0910)	(2,201.0700)
Wall5:Stone	4,667.9450	5,055.6290	5,283.5430	5,351.8710
	(3,989.9740)	(3,933.8350)	(3,907.5580)	(3,907.2050)
Wall6:Wood	-2,980.3520	-2,146.6110	-1,809.5930	-1,797.0390
	(2,197.0210)	(2,166.2940)	(2,151.8900)	(2,151.8880)
Wall7:Mixed Material	2,059.9460	5,230.8420*	5,598.5640*	5,566.8690*
	(2,229.8740)	(2,201.6200)	(2,186.9940)	(2,186.8650)
Number of Bedroom	-6,393.1400***	-4,331.6180***	-4,228.5450***	-4,240.5540***
	(364.0816)	(367.6718)	(365.2565)	(365.1302)
Number of Bathroom	13,182.1100***	1,210.0520	3,202.4520*	3,186.1720*
	(583.8316)	(1,619.5630)	(1,612.3290)	(1,612.2920)
Number of Bathroom2		2,882.7340***	2,257.3610***	2,259.6090***
		(400.4479)	(399.2013)	(399.2013)
Frontage	51.5925***	64.4182***	62.9407***	67.2671***
	(6.9885)	(6.9070)	(6.8613)	(5.8955)
Garage1:Basement	11,541.0600**	13,715.4400***	13,577.7900***	13,508.6700***
	(3,610.8350)	(3,561.2690)	(3,537.4710)	(3,537.0630)
Garage2:Attached	9,810.5470***	12,365.5700***	12,534.2900***	12,428.6100***
	(939.8244)	(931.4330)	(925.2516)	(921.2801)
Garage3:Detached	7,415.7560***	7,699.1240***	7,694.9020***	7,653.5440***
	(855.9870)	(843.9951)	(838.3533)	(837.6901)
Garage4:Carport	3,377.8640	2,272.7690	1,983.0060	1,919.8150
	(2,203.4550)	(2,172.8610)	(2,158.3930)	(2,157.8060)

Garage Area(square footage)	13.4007***	13.7832***	13.9404***	14.0184***
	(1.5285)	(1.5072)	(1.4972)	(1.4959)
Lot Size(square feet)	0.0168	0.0116	0.0112	
	(0.0093)	(0.0091)	(0.0091)	
Distance	1,439.0130***	1,495.6800***	1,515.4030***	1,531.7590***
	(69.3472)	(68.4063)	(67.9574)	(66.6497)
Quarter2	4,009.0790***	4,111.2270***	4,003.5140***	3,998.4810***
	(584.4816)	(576.2632)	(572.4407)	(572.4320)
Quarter3	4,659.1330***	4,799.7250***	4,687.3380***	4,688.1580***
	(581.5688)	(573.4071)	(569.6064)	(569.6119)
Quarter4	2,580.5030***	2,739.2490***	2,642.1940***	2,643.4380***
	(628.2356)	(619.4203)	(615.3021)	(615.3075)
AIC	596429.94	595712.5	595348.88	595348.41
BIC	596625.32	595924.16	595560.54	595551.92
Observations	25,357	25,357	25,356	25,356
R^2	0.7305	0.7380	0.7415	0.7415
Adjusted R^2	0.7302	0.7378	0.7413	0.7413
Residual Std. Error	30,983.9700 (df = 25334)	30,547.5300 (df = 25332)	30,343.3400 (df = 25331)	30,343.6500 (df = 25332)
F Statistic	3,120.9120*** (df = 22; 25334)	2,973.6290*** (df = 24; 25332)	3,028.0790*** (df = 24; 25331)	3,159.6040*** (df = 23; 25332)
Note:				*p**p***p<0.001

Table 2 Regression Results

(e)

1. Year the house built has a positive impact on house price, comparing with the house built in the past, price of the house built one year later will increase by approximately 466.
2. Stories increased by one unit would lead to an increase of 1854 in house price.
3. The Larger the living area, the higher the selling price is. One square footage will lead to a rise of 16 in price.
4. Wall is a categorical variable. Each type of material used for wall construction is controlled by dummy variable 0 and 1. 1 means use that type of material while 0 means do not use that type of material. Based on this, for each house, model can control only that type of material used have impact on house price.

   Different Wall materials have different influences on price. Type 2 and 6 have a negative impact on price while types 3, 4, 5 have a positive effect. Whether to use one of these types of materials would lead to the fluctuation in price from 972 to 10573.
5. The number of bedrooms has a huge influence on house price, the fewer bedrooms, the much lower the house price. Reducing one room will result in a price reduction of 5000.
6. The number of bathrooms has opposite trends towards price in comparison with

7. Frontage of house has relatively little impact on house price, it is a positive influence.
8. Type of garages is also a categorical variable. Different garage types would lead to varying degrees of price growth. For each kind of garage, dummy variables control the impact of garage towards house price. Noticeably, basement and attached garage have significantly positive impact on house price while the other two types of garage only lead to relatively slow increase in the house price.
9. The larger the garage is, the higher the house price is. But this effect is small.
10. The quarter the house sold has positive effect on the price. Sale Date lies in quarter 2 and quarter 3 have more larger impact on price than quarter 4. The price of buying a house in different seasons is about 1000.

(f)

It is interesting that the number of bedrooms have negative impact on the price while the number of bathrooms have positive influence on house price.

Using some kinds of wall material would lead to an increase on house price such as stone while some would have reduction of house price like wood. This may because house built by wood is not as strong as house constructed by wood to resist in different weather. Wood house is not welcomed by most consumers, different demand relationships have caused them to have different effects on housing prices.

Appendix

```
library(haven)
library(ggplot2)
library(lmtest) # linear model tests, like tests for heteroskedasticity, reset test, etc.
library(sandwich) # to use HC standard errors, if needed
library(car)# VIF test, etc.
library(stargazer)
house <- read_dta("Desktop/R applied econometrics/quiz/Midterm/house.dta")
View(house)
house$wall<- as.factor(house$wall)
house$garage_type <- as.factor(house$garage_type)
house$garage <- as.factor(house$garage)
house$sold93 <-as.factor(house$sold93)
house$sold94 <-as.factor(house$sold94)
house$sold95 <-as.factor(house$sold95)
house$sold96 <-as.factor(house$sold96)
house$sold97 <-as.factor(house$sold97)
house$sold98 <-as.factor(house$sold98)
###1.Data Inspection and Statistical Inference
#a
View(house)
dim(house)
str(house)
summary(house)
#b
```

```
attach(house)
options(scipen=100)
hist(price,xlab='Price',main='Histogram of Price')
#c
cor(sqft,price)
anova(lm(price ~    garage_type))
plot(garage_type,price,xlab='Garage Type',ylab='Price',main='Garage Type versus Price')
#d
hist(saledate,breaks = 'months',freq=T,xlab='Sale Date',col='grey')
#hist(saledate,breaks = 'years',freq=T)
#hist(saledate,breaks = 'weeks',,freq=T)
#hist(saledate,breaks = 'days',,freq=T)

#2.Model Selection and Output
cor(price,yrbuilt)
anova(lm(price ~    stories))
anova(lm(price ~    wall))
cor(price,sqft)
cor(price,bed)
cor(price,bath)
cor(price,halfbath)
cor(price,fullbath)
cor(price,frontage)
cor(price,depth)
anova(lm(price ~ garage_type))
cor(price,garage_sqft)
t.test(price ~ garage)
cor(price,rooms)
cor(price,lotsize)
cor(price,distance)
anova(lm(price ~ saleq))

par(mfrow=c(2,2))
plot(yrbuilt,price,cex=0.5,pch=20)#
plot(stories,price)#
plot(sqft,price,cex=0.5,pch=20)#
plot(wall,price)
plot(bed,price)#
plot(bath,price)#
plot(fullbath,price)
plot(halfbath,price)
plot(frontage,price)#
plot(depth,price)
plot(garage_type,price)#
```

```
plot(rooms,price)
plot(lotsize,price)#
plot(longitude,price)
plot(latitude,price)
plot(saledate,price)
plot(distance,price)
#create a new variable named distance
house$n1 <- abs((-(house$longitude)-83.5379))
house$n2 <- (abs(house$latitude-41.6528))/40
house$distance <- sqrt((house$n1*69)^2 + (house$n2*53)^2)
View(house)
attach(house)
#date
house$saleq <- quarters(house$saledate)
#bed bath rooms ->choose bed bath
mod1 <- lm(price ~ yrbuilt + stories + sqft + wall + bed + bath + frontage +
                    garage_type + garage_sqft + lotsize + distance + saleq)#better one
summary(mod1)
plot(mod1)
mod1.1 <- lm(price ~ yrbuilt + stories + sqft + wall + rooms + frontage +
                       garage_type + garage_sqft + lotsize + distance + saleq)#not rooms
summary(mod1.1)
plot(mod1.1)
stargazer(mod1, mod1.1,
                add.lines = list(
                   c("AIC", round(AIC(mod1),4), round(AIC(mod1.1),4)),
                   c("BIC", round(BIC(mod1),4), round(BIC(mod1.1),4))
                ),
                type="text", no.space=TRUE) # bed bath
#garage garage_type ->choose garage_type
mod1.2 <- lm(price ~ yrbuilt + stories + sqft + wall + bed + bath + frontage +
                     garage + garage_sqft + lotsize + distance + saleq)
summary(mod1.2)
plot(mod1.2)
stargazer(mod1, mod1.2,
                add.lines = list(
                   c("AIC", round(AIC(mod1),4), round(AIC(mod1.2),4)),
                   c("BIC", round(BIC(mod1),4), round(BIC(mod1.2),4))
                ),
                type="text", no.space=TRUE) #mod1 is better
#mod1
e <- resid(mod1)
summary(e)
plot(mod_3)
```

```
plot(yrbuilt,price)
abline(lm(price ~ yrbuilt),col='red')
vif(mod1)# no high collinearity among x's

#should higher powers be included in regression?
#heter....
par(mfrow=c(2,2))
resettest(mod1,power=3) # reject at least one X's square term's coefficient is statistically significant
plot(yrbuilt,resid(mod1))# less
abline(h=0,col='red')
plot(stories,resid(mod1))
abline(h=0,col='red')
plot(wall,resid(mod1))
abline(h=0,col='red')
plot(sqft,resid(mod1),xlab='Area of Living Area(square footage)',
     ylab ='Residuals of Modeal 1',main='Living Area & residuals')#need square
abline(h=0,col='red')
plot(bed,resid(mod1))
abline(h=0,col='red')
plot(bath,resid(mod1),xlab= 'Number of Bathrooms', ylab ='Residuals of Modeal 1',main='Bathrooms & residuals ')#need square
abline(h=0,col='red')
plot(frontage,resid(mod1))
abline(h=0,col='red')
plot(garage_sqft,resid(mod1))
plot(lotsize,resid(mod1))
plot(distance,resid(mod1))
abline(h=0,col='red')

# add square terms to sqft and bath
mod2 <- lm(price ~ yrbuilt + stories + sqft + I(sqft^2) + wall + bed + bath+ I(bath^2) + frontage +
             garage_type + garage_sqft + lotsize + distance + saleq)
summary(mod2)
plot(mod2)
par(mfrow=c(1,1))#influential outlier:16017
stargazer(mod1, mod2,
          add.lines = list(
            c("AIC", round(AIC(mod1),4), round(AIC(mod2),4)),
            c("BIC", round(BIC(mod1),4), round(BIC(mod2),4))
          ),
          type="text", no.space=TRUE)
mod2.1<- lm(price ~ yrbuilt + stories + sqft + I(sqft^2) + wall + bed + bath+ I(bath^2) + frontage +
              garage_type + garage_sqft + lotsize + distance + saleq,data=house[-16017,])
summary(mod2.1)
```

```
stargazer(mod1, mod2,mod2.1,type="text", no.space = TRUE)
#still suspect heteroskedasticity
bptest(mod2.1) # reject H0:constant variance
#heteroskedasticity-corrected standard error
HCcov2.1 <- vcovHC(mod2.1, type="HC1")
rse2.1 <- sqrt(diag(HCcov2.1))
coeftest(mod2.1, vcov=HCcov2.1)
stargazer(mod2.1, mod2.1,
          se=list(NULL, rse2.1),
          column.labels = c("Mod 2.1", "2.1 HC"),
          no.space = TRUE,
          type = "text")
linearHypothesis(mod2.1, c('lotsize'), vcov=HCcov2.1) # not right model remove lotsize
mod2.2<- lm(price ~ yrbuilt + stories + sqft + I(sqft^2) + wall + bed + bath+ I(bath^2) + frontage +
              garage_type + garage_sqft + distance + saleq,data=house[-16017,])#mod2.2 is
best
summary(mod2.2)
bptest(mod2.2)
HCcov2.2 <- vcovHC(mod2.2, type="HC1")
rse2.2 <- sqrt(diag(HCcov2.2))
coeftest(mod2.2, vcov=HCcov2.2)
stargazer(mod2.2, mod2.2,
          se=list(NULL, rse2.2),
          column.labels = c("Mod 2.2", "2.2 HC"),
          no.space = TRUE,
          type = "text")
#log
mod3<- lm(log(price) ~ yrbuilt + stories + sqft + I(sqft^2) + wall + bed + bath+ I(bath^2) + frontage
+
            garage_type + garage_sqft + distance + saleq,data=house[-16017,])
summary(mod3) #bad
```

```
#model compare
stargazer(mod1, mod2, mod2.1, mod2.2,
          no.space = TRUE, digits=4, star.cutoffs = c(0.05, 0.01, 0.001),
          add.lines=list(
                          c("AIC", round(AIC(mod1),2), round(AIC(mod2),2),
                            round(AIC(mod2.1),2), round(AIC(mod2.2),2)),
```

```
                                c("BIC", round(BIC(mod1),2), round(BIC(mod2),2),
                                  round(BIC(mod2.1),2), round(BIC(mod2.2),2))),
          covariate.labels= c("Year Built", "Stories","Living Area(square footage)","Living Area(square footage)^2",
                              "Wall2:Concrete block or tile","Wall3:Aluminum,vinyl,or steel siding","Wall4:Brick",
                              "Wall5:Stone","Wall6:Wood","Wall7:Mixed Material","Number of Bedroom",
                              "Number of Bathroom","Number of Bathroom^2","Frontage","Garage1:Basement","Garage2:Attached",
                              "Garage3:Detached","Garage4:Carport","Garage Area(square footage)","Lot Size(square feet)",
                              "Distance","Quarter2","Quarter3","Quarter4"),
          dep.var.labels = "Sale Price",
          omit = c("Constant"),
          type="html", out="~/Desktop/table-1.html")
stargazer(mod1, mod2, mod2.1, mod2.2,
          no.space = TRUE, digits=4, star.cutoffs = c(0.05, 0.01, 0.001),
          add.lines=list(
            c("AIC", round(AIC(mod1),2), round(AIC(mod2),2),
              round(AIC(mod2.1),2), round(AIC(mod2.2),2)),
            c("BIC", round(BIC(mod1),2), round(BIC(mod2),2),
              round(BIC(mod2.1),2), round(BIC(mod2.2),2))),
          covariate.labels= c("Year Built", "Stories","Living Area(square footage)","Living Area(square footage)2",
                              "Wall2:Concrete block or tile","Wall3:Aluminum,vinyl,or steel siding","Wall4:Brick",
                              "Wall5:Stone","Wall6:Wood","Wall7:Mixed Material","Number of Bedroom",
                              "Number of Bathroom","Number of Bathroom2","Frontage","Garage1:Basement","Garage2:Attached",
                              "Garage3:Detached","Garage4:Carport","Garage Area(square footage)","Lot Size(square feet)",
                              "Distance","Quarter2","Quarter3","Quarter4"),
          dep.var.labels = "House Price",
          omit = c("Constant"),
          type="html", out="~/Desktop/table-1.html")


###Final Model
mod2.2<- lm(price ~ yrbuilt + stories + sqft + I(sqft^2) + wall + bed + bath+ I(bath^2) + frontage +
              garage_type + garage_sqft + distance + saleq,data=house[-16017,])#mod2.2 is best
summary(mod2.2)
```