PacDev Case Module 6

Ali Ladha

2024-11-01

- Q1
- Q2
- Q3
- Q4
- Q5

```
library(tidyverse)

sfr <- read.csv('/Users/aliladha/Documents/Files/School Work/College/Graduate/IS 6489-00
1 Stats & Pred Analytics/pacdev_data.csv')

sfr <- sfr %>% mutate(city = factor(city))

glimpse(sfr)
```

```
## Rows: 4,991
## Columns: 5
## $ city  <fct> Santa Monica, Long Beach, Westwood, Westwood, Santa Monica, West…
## $ sqft  <int> 2343, 1705, 1846, 1782, 2257, 2098, 1838, 1907, 1842, 1474, 1764…
## $ bed   <int> 1, 1, 3, 2, 1, 4, 2, 4, 2, 2, 1, 3, 3, 2, 5, 6, 3, 3, 3, 5, 3, 2…
## $ bath  <int> 1, 1, 2, 1, 1, 3, 1, 3, 1, 1, 1, 2, 2, 1, 4, 5, 2, 2, 2, 4, 2, 1…
## $ price <int> 537227, 432815, 495542, 151566, 784151, 369506, 456937, 503307, …
```

```
str(sfr)
```

```
## 'data.frame':    4991 obs. of  5 variables:
##  $ city : Factor w/ 3 levels "Long Beach","Santa Monica",..: 2 1 3 3 2 3 1 1 1 3 ...
##  $ sqft : int  2343 1705 1846 1782 2257 2098 1838 1907 1842 1474 ...
##  $ bed  : int  1 1 3 2 1 4 2 4 2 2 ...
##  $ bath : int  1 1 2 1 1 3 1 3 1 1 ...
##  $ price: int  537227 432815 495542 151566 784151 369506 456937 503307 485277 650842
...
```

Q1

```
#1.  Fit a simple linear regression model of price ~ sqft. Interpret the coefficients fo
r the intercept and sqft.
# To interpret model coefficients in this case study means: write down in concrete terms
what each coefficient says about the value of a home, or the change in the value of a ho
me, conditional on predictors. For example, suppose the coefficient for x is 1.5. To int
erpret this number would entail saying something like this: “An increase in 1 in x is as
sociated with a change of 1.5 in y, on average, holding the other predictors constant.”

(sfr_lm <- lm(price ~ sqft, data = sfr))
```

```
##
## Call:
## lm(formula = price ~ sqft, data = sfr)
##
## Coefficients:
## (Intercept)         sqft
##     40623.0        269.3
```

```
summary(sfr_lm)
```

```
##
## Call:
## lm(formula = price ~ sqft, data = sfr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -622948 -151283   -1650  138951  804553
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40623.019  15862.454   2.561   0.0105 *
## sqft          269.345      7.742  34.791   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 210300 on 4989 degrees of freedom
## Multiple R-squared:  0.1953, Adjusted R-squared:  0.1951
## F-statistic:  1210 on 1 and 4989 DF,  p-value: < 2.2e-16
```

Intercept: if the square footage is 0 (x), the average price of the home (y) is $40,623 holding the other predictors constant if applicable (in this model so far there are no other predictors for now) . This is impossible since homes must have square footage to be considered a home

Sqft: for every 1 unit increase in square footage (x), the average price of the home (y) increases by $269.30.

```
#2 Center sqft and refit the model. The intercept changes a lot after centering. Why?
# Note: there are several options for centering.
#
# Create a new centered variable in your data and use that in the model. You would simply subtract the mean of the column from every observation.
# Use the scale() function in the model formula: scale(sqft, scale = F). (scale = F means: center but don’t scale. The default for the function is to center and scale.)
# Do the centering yourself in the model formula using the I() function: I(sqft - mean(sqft)).


#using the Scale method for centering:

sfr_lm_center <- lm(price ~ scale(sqft, scale = F), data = sfr)
sfr_lm_center
```

```
## 
## Call:
## lm(formula = price ~ scale(sqft, scale = F), data = sfr)
## 
## Coefficients:
##            (Intercept)  scale(sqft, scale = F)  
##               582698.7                   269.3
```

```
summary(sfr_lm_center)
```

```
## 
## Call:
## lm(formula = price ~ scale(sqft, scale = F), data = sfr)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -622948 -151283   -1650  138951  804553 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            5.827e+05  2.976e+03  195.77   <2e-16 ***
## scale(sqft, scale = F) 2.693e+02  7.742e+00   34.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 210300 on 4989 degrees of freedom
## Multiple R-squared:  0.1953, Adjusted R-squared:  0.1951 
## F-statistic:  1210 on 1 and 4989 DF,  p-value: < 2.2e-16
```

After centering the intercept is interpreted as the Average price of the Single Family Residences/homes when the square footage is Average (instead of 0) which is $582,698.7. This is why the intercept is much higher compared to the non centered intercept which is based on the square footage being 0.

The increase in price from 1 unit of square footage is still the same at $269.3.

The p value for square footage is <0.05 which means that the square footage is statistically significant.

Q2

```
#Fit a multiple regression model of price using all the available predictors. Center the
numeric predictors. Interpret the coefficients in this model (including the intercept).
#Remember that Long Beach is the (missing) reference city (assuming that factor levels h
ave been assigned alphabetically).



(sfr_m <- lm(price ~ scale(sqft, scale = F) + city + scale(bed, scale = F) + scale(bath,
scale = F), data = sfr))
```

```
##
## Call:
## lm(formula = price ~ scale(sqft, scale = F) + city + scale(bed,
##     scale = F) + scale(bath, scale = F), data = sfr)
##
## Coefficients:
##            (Intercept)  scale(sqft, scale = F)        citySanta Monica
##               513686.8                   272.0                189818.8
##           cityWestwood   scale(bed, scale = F)  scale(bath, scale = F)
##                88101.3                  -524.5                 -2471.7
```

```
summary(sfr_m)
```

```
##
## Call:
## lm(formula = price ~ scale(sqft, scale = F) + city + scale(bed,
##     scale = F) + scale(bath, scale = F), data = sfr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -546425 -138158   -2262  124719  845241
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            513686.845   3887.942 132.123   <2e-16 ***
## scale(sqft, scale = F)    272.038      7.516  36.193   <2e-16 ***
## citySanta Monica       189818.758   6759.691  28.081   <2e-16 ***
## cityWestwood            88101.333   6798.754  12.958   <2e-16 ***
## scale(bed, scale = F)    -524.490   8418.596  -0.062    0.950
## scale(bath, scale = F)  -2471.708   9440.014  -0.262    0.793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 195100 on 4985 degrees of freedom
## Multiple R-squared:  0.3074, Adjusted R-squared:  0.3067
## F-statistic: 442.6 on 5 and 4985 DF,  p-value: < 2.2e-16
```

```
summary(factor(sfr$city))
```

```
##   Long Beach Santa Monica     Westwood 
##         2520         1246         1225
```

The intercept is the Average price of a Single Family Residence in Long Beach $513,686.80 when square footage is average when all predictors are 0, or categorical are at the reference levels. sqft: the average increase in price being $272.0 for every 1 unit increase of square footage regardless of the city. citySanta Monica: the average price increases by approximately: $189,818.80 if the city is Santa Monica compared to Long Beach. cityWestwood: the average price increases by approximately: $88,101.33 if the city is Westwood compared to Long Beach bed: the average decrease in price by $-524.50 per addition of bedroom holding all other factors constant bath: the average decrease in -$2471.7 price per addition of bathroom holding all other factors constant

The p value for square footage, city for santa monica and the city for westwood is below 0.05. Therefore those variables are statistically significant.

Q3

```
# To the above model add an interaction between centered sqft and city. This means that
you combine these terms multiplicatively (*) rather than additively (+).
#
# Create a visualization of the interaction between sqft and city.
#
# Interpret the two interaction coefficients in this model.
#
# Interaction models can be tricky to understand. Here is some guidance:

#
# The intercept is the average value of the target when the inputs are 0 (for numeric va
riables) or the reference category (for categorical variables).
#
# The main effects are the non-interaction coefficients in the output. In the plot you c
reated for this question you can see that there is a regression line for each city. Simi
larly in the interaction model: there is no single relationship between price and sqft,
and consequently the main effect for sqft is conditional on city. Specifically, it denot
es the relationship between sqft and price for the reference city. The same is true for
the main effects ofcity: they are conditional on sqft = 0. The main effects for a predic
tor in an interaction model will always be conditional on the levels of the variable wit
h which it has been interacted.
#
# The interaction effects are the final 2 coefficients in the output. (The colon indicat
es the interaction, as in sqft:citySanta Monica.) These coefficients estimate the change
in the slope of the regression line for each city compared to the reference city. If the
interaction coefficients are positive that means that the regression line relating sqft
to price is steeper for that particular city in comparison to the reference city, or, eq
uivalently, that the relationship is stronger.


(sfr_m_interaction <- lm(price ~ scale(sqft, scale = F) * city + scale(bed, scale = F) +
scale(bath, scale = F), data = sfr))
```

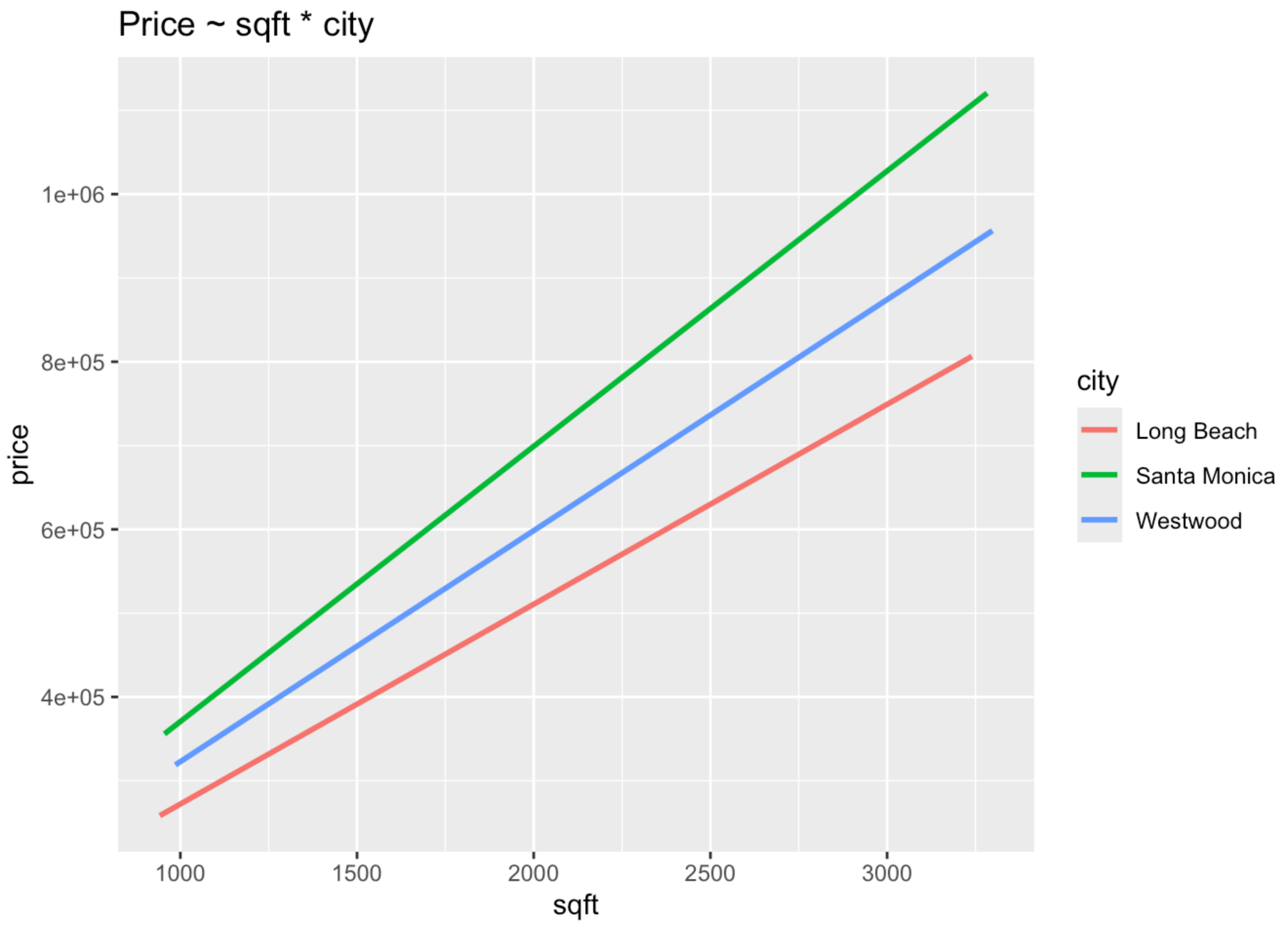
```
##
## Call:
## lm(formula = price ~ scale(sqft, scale = F) * city + scale(bed,
##     scale = F) + scale(bath, scale = F), data = sfr)
##
## Coefficients:
##                             (Intercept)
##                               513735.00
##                  scale(sqft, scale = F)
##                                  241.52
##                        citySanta Monica
##                               189480.25
##                            cityWestwood
##                                88114.56
##                   scale(bed, scale = F)
##                                 -338.11
##                  scale(bath, scale = F)
##                                -2468.87
## scale(sqft, scale = F):citySanta Monica
##                                   89.63
##     scale(sqft, scale = F):cityWestwood
##                                   36.65
```

```
summary(sfr_m_interaction)
```

```
## 
## Call:
## lm(formula = price ~ scale(sqft, scale = F) * city + scale(bed, 
##     scale = F) + scale(bath, scale = F), data = sfr)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -543099 -138790   -3007  126348  861025 
## 
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             513735.00    3878.45 132.459  < 2e-16
## scale(sqft, scale = F)                     241.52      10.14  23.812  < 2e-16
## citySanta Monica                        189480.25    6743.50  28.098  < 2e-16
## cityWestwood                             88114.56    6783.00  12.991  < 2e-16
## scale(bed, scale = F)                     -338.11    8401.82  -0.040   0.9679
## scale(bath, scale = F)                   -2468.87    9419.85  -0.262   0.7933
## scale(sqft, scale = F):citySanta Monica     89.63      17.49   5.124 3.11e-07
## scale(sqft, scale = F):cityWestwood         36.65      18.05   2.030   0.0424
## 
## (Intercept)                             ***
## scale(sqft, scale = F)                  ***
## citySanta Monica                        ***
## cityWestwood                            ***
## scale(bed, scale = F)                      
## scale(bath, scale = F)                     
## scale(sqft, scale = F):citySanta Monica ***
## scale(sqft, scale = F):cityWestwood     *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 194700 on 4983 degrees of freedom
## Multiple R-squared:  0.3111, Adjusted R-squared:  0.3101 
## F-statistic: 321.5 on 7 and 4983 DF,  p-value: < 2.2e-16
```

```
sfr %>% ggplot(aes(x = sqft, y = price, col = city)) + geom_smooth(method = "lm", se =
F) + labs(title = 'Price ~ sqft * city')
```

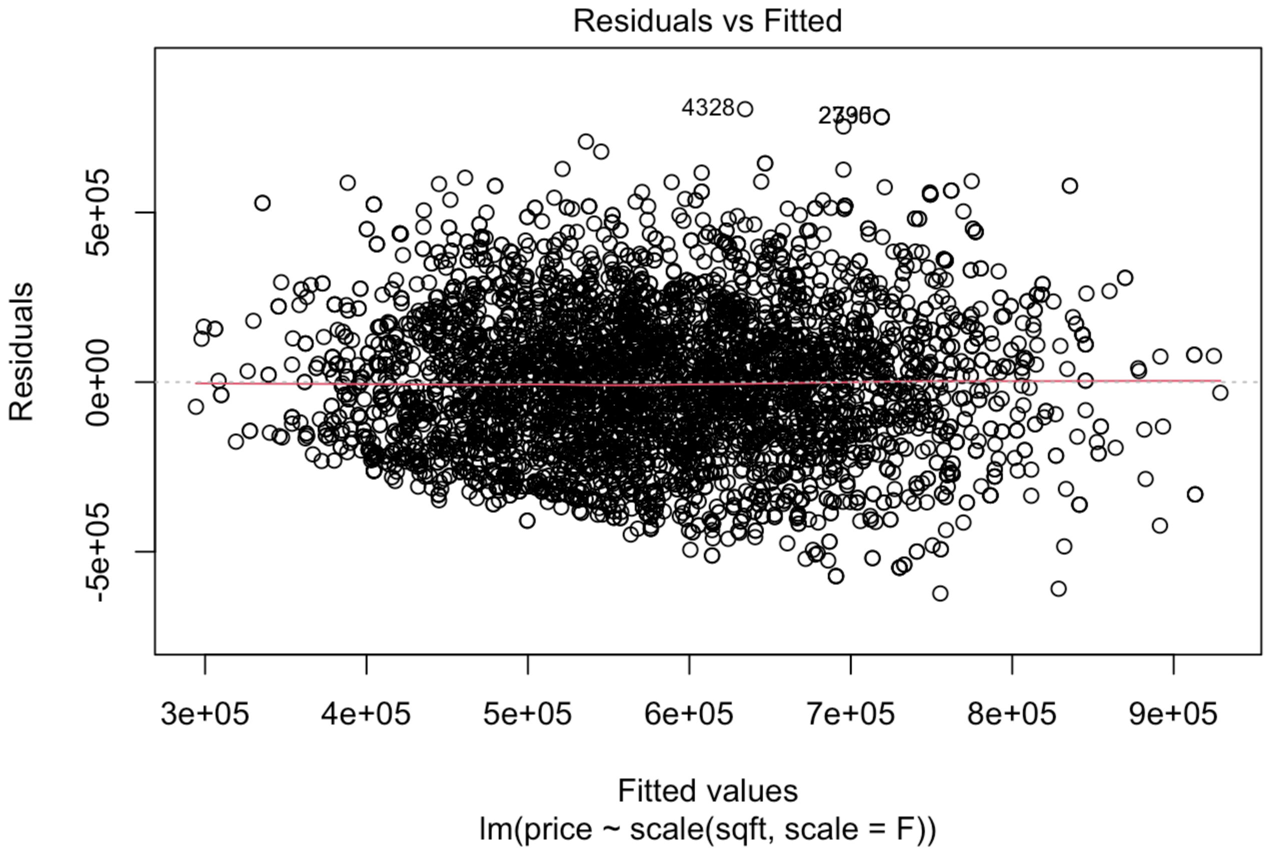
```
## `geom_smooth()` using formula = 'y ~ x'
```



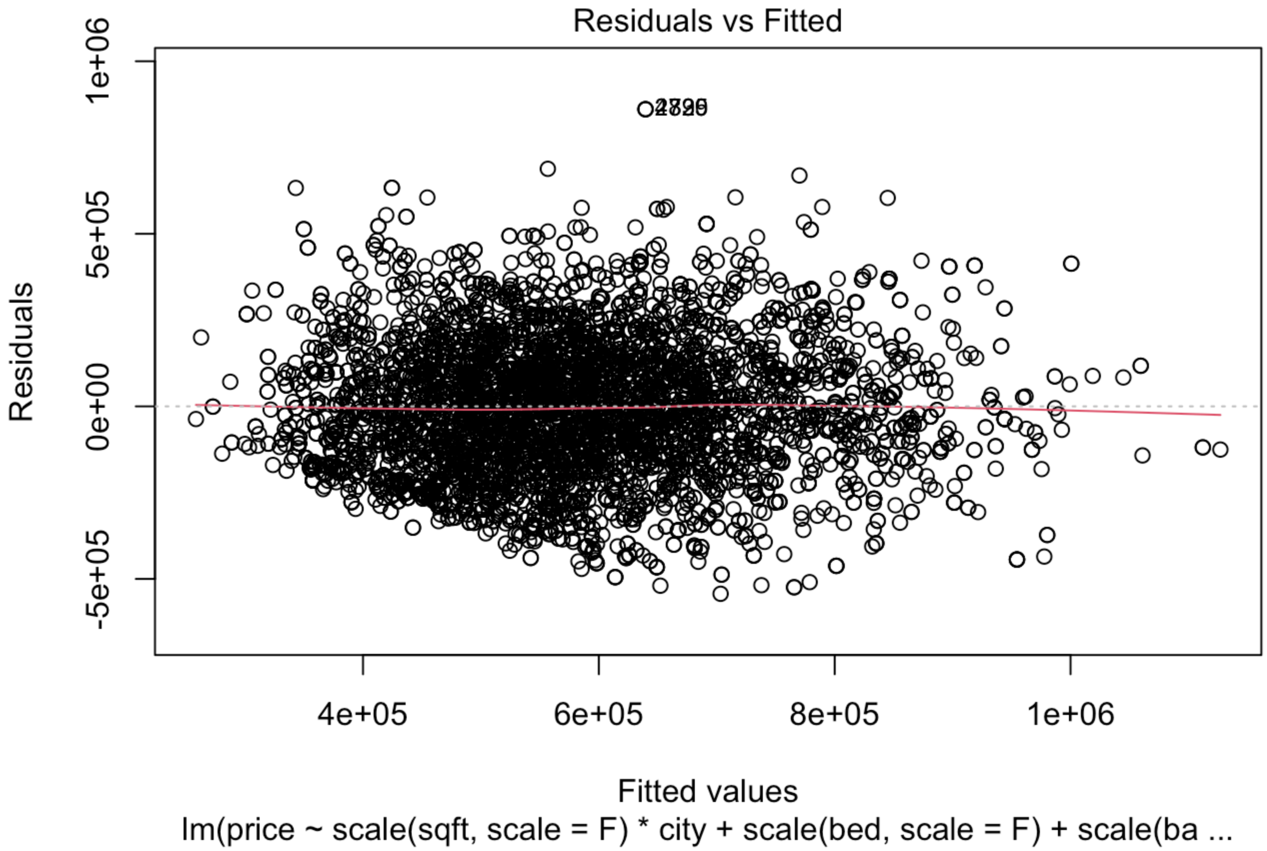
sqft:citySanta Monica - The average increase in price of $89.63 for every 1 unit addition in square footage in Santa Monica compared to Long Beach holding all the other predictors constant. sqft:cityWestwood - The average increase in price of $36.65 for every 1 unit addition in squarefootage in Westwood compared to Long Beach holding all the other predictors constant.

Q4

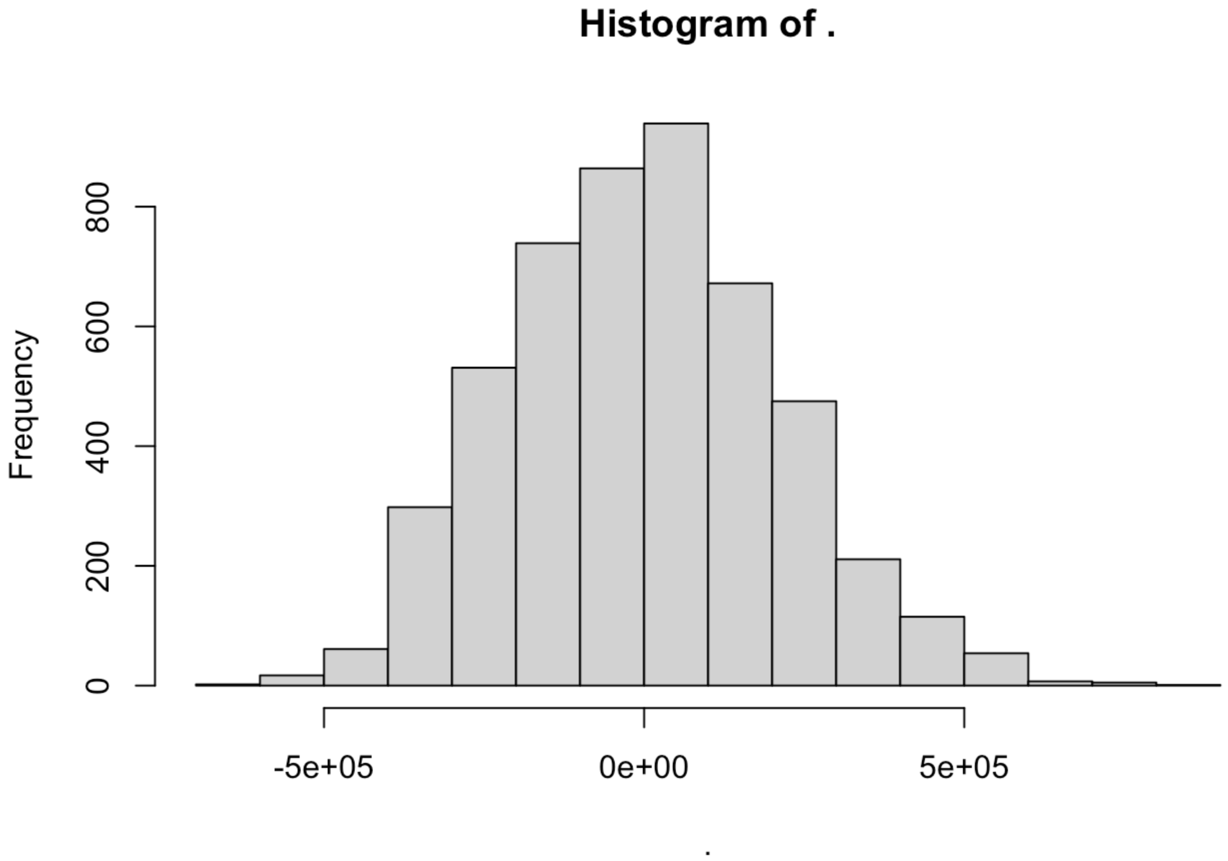
```
# Is this a good model? To assess the interaction model fit create a residual plot with
model residuals on the vertical axis and the fitted values on the horizontal axis. One e
asy way to get a residual plot in R is to use plot(model, which = 1). Comment on model f
it.
#
#


#comparing multiple regression model Without the interaction
plot(sfr_lm_center, which = 1)
```



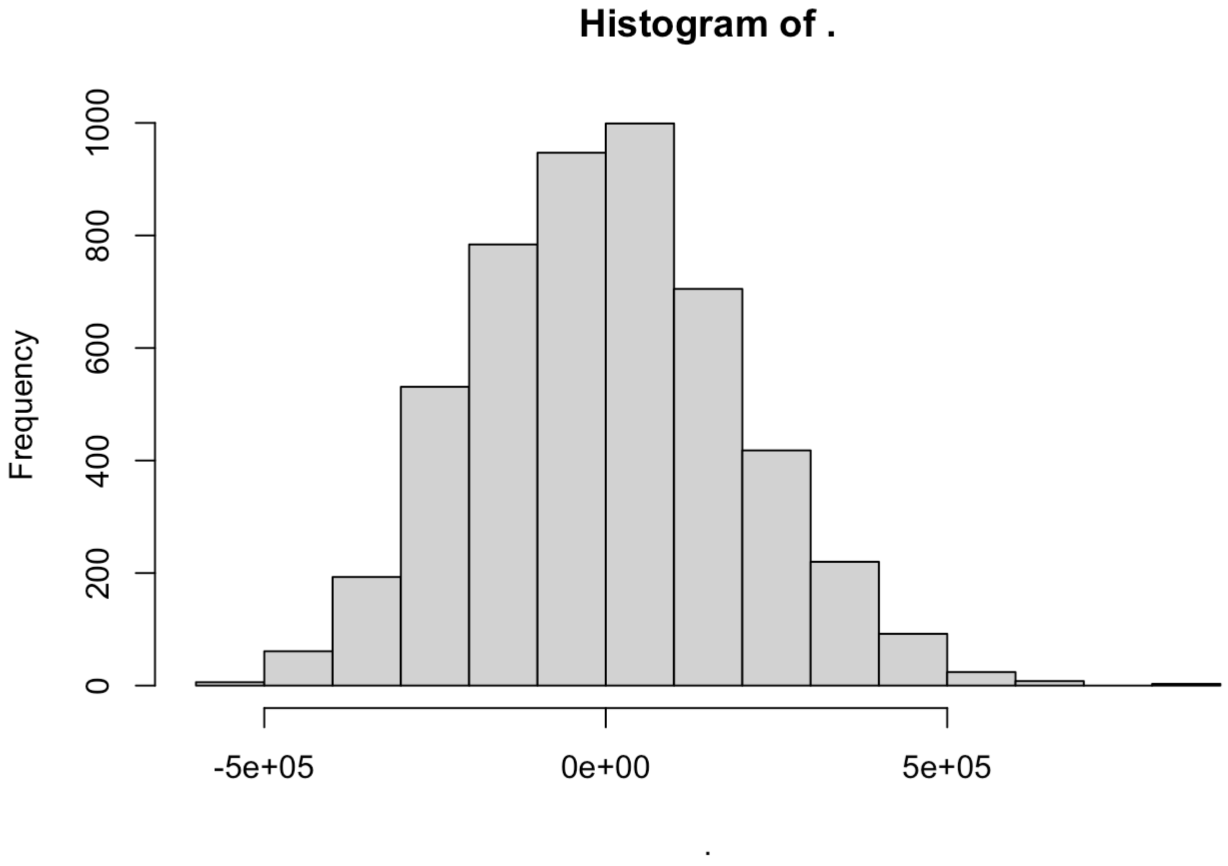
```
#comparing multiple regression model With the Interaction
plot(sfr_m_interaction, which = 1)
```



```
#comparing multiple regression model Without the interaction in a Histogram
sfr_lm_center %>% resid() %>% hist()
```



```
#comparing multiple regression model With the Interaction in a Histogram
sfr_m_interaction %>% resid() %>% hist()
```



According to the histogram, the interaction model is a good fit as the residuals are normally distributed. There are outliers though which is visible in the Residual plot which shows variability in the data. It may affect the accuracy of the model too. Most of the points are at 0, but there is slightly more points below the mean (0) compared to the number of points above the mean.

Secondly the data points of the residual plots before the histogram is not funneled therefore there is no/minimal heteroskedastic errors. It is more of a fan shape. Therefore, the model is still considered to be a good fit.

Q5

```
#What should Andrew say in his presentation? Write a brief summary of the quantitative e
vidence that he should use to support this recommendation.
```

During his presentation, Andrew should mention that PacDev should focus on prioritizing work in Santa Monica.

When the square footage of the single family residences is average, the average price of the single family residence is: $513,735.00 in Long Beach. Santa Monica has an additional higher single family residence price of $189,480.25 compared to Long Beach. Whereas, Westwood has an additional higher single family residence price $88,114.56.

Furthermore, Comparing Santa Monica to Long Beach: the increase in price for each addition in sq footage in Santa Monica is: $89.63. Whereas when comparing Westwood to Long Beach: the increase in price for each addition in sq footage in Westwood is: $36.65. Bedrooms decrease the price by: $338.11 on average regardless of the city Bathrooms decrease the price by: $2468.87 on average regardless of the city.

The interaction plot is also good to mention to PacDev since it shows its a good model to prove the data is accurate. So PacDev should remodel and sell homes in Santa Monica while building fewer bedrooms and bathrooms. Of course a home must have atleast a bathroom and preferably a bedroom, but these should be limited to maximize profit as they decrease the average price of the single family residence.