Assignment 1

Earle Aguilar

April 5, 2018

Problem 1

a)

Inference Questions:

Does the location affect the price of real state?
Does the number of baths affect the price of real state?

Prediction Questions:

Do larger square feet increase the price of real state?
Does the type of real state increase the price?

a)

Does the location affect the price of real state?

```r
df <- read.csv("~/GitHub/LArealestate.csv")
summary(df)
```

```
##                        address         beds            baths
##                            :  1   Min.   : 1.000   Min.   :   1.00
##    1005 Benedict Canyon Dr:  1   1st Qu.: 3.000   1st Qu.:   2.00
##    10084 Westwanda Dr     :  1   Median : 4.000   Median :   3.75
##    1009 N Beverly Dr      :  1   Mean   : 3.902   Mean   :  15.32
##    1010 N Rexford Dr      :  1   3rd Qu.: 5.000   3rd Qu.:   6.00
##    10101 Angelo View Dr   :  1   Max.   :10.000   Max.   :2822.00
##  (Other)                  :249   NA's   :1        NA's   :1
##       sqft
##  Min.   :  548
##  1st Qu.: 1484
##  Median : 2987
##  Mean   : 3897
##  3rd Qu.: 5285
##  Max.   :29000
##  NA's   :1
##                                                                              date
##  03/24/2014Coldwell Banker Residential Brokerage - Beverly Hills NorthFeatured : 19
##  03/24/2014Coldwell Banker Residential Brokerage - Beverly Hills SouthFeatured :  9
##  03/24/2014Sotheby's International Realty -Featured                            :  9
##  02/25/2014Hilton & Hyland                                                     :  7
##  03/24/2014Rodeo Realty - Beverly Hills                                        :  7
##  03/24/2014Rodeo Realty Inc.                                                   :  5
##  (Other)                                                                       :199
##      price                      city         type
##  Min.   :    2195   Beverly Hills:148   condo: 39
```

```
##  1st Qu.:  762500   culver city  : 30   sfh  :216
##  Median : 2200000   Culver City  : 28
##  Mean   : 4388878   Palms        : 49
##  3rd Qu.: 5542500
##  Max.   :43000000
##
```

The city column has double culver city entries with different capilatization. Im going to create a new column ('city2') which has the correct number of factors.

```r
city2 <- as.factor(tolower(as.character(df$city)) )
df$city2 <- city2
```

First I'm doing a preliminiary check to determine how significant the city is in the model in comparison to other variables of interest.

```r
summary(lm(df$price ~ df$city2 + df$beds + df$baths + df$sqft + df$type, df))
```

```
##
## Call:
## lm(formula = df$price ~ df$city2 + df$beds + df$baths + df$sqft +
##     df$type, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7633726 -1226366  -210354   478163 21046043
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           279891.2   802325.7   0.349   0.7275
## df$city2culver city -1241409.1   611893.9  -2.029   0.0436 *
## df$city2palms       -1236582.2   578733.9  -2.137   0.0336 *
## df$beds               154703.9   247992.0   0.624   0.5333
## df$baths             -263161.3   218968.4  -1.202   0.2306
## df$sqft                 1461.3      139.6  10.469   <2e-16 ***
## df$typesfh           -648569.8   681986.8  -0.951   0.3425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2989000 on 246 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.7298, Adjusted R-squared:  0.7232
## F-statistic: 110.7 on 6 and 246 DF,  p-value: < 2.2e-16
```

The above output shows that aside from square feet the next significant parameter is the city so I will make a model with just real state price and the city.

Fitting Model

```r
model <- lm(df$price ~ df$city2, df)
summary(model)
```

```
##
## Call:
## lm(formula = df$price ~ df$city2, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -6121145 -3108645  -252061   593033 36003855
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          6996145     392465  17.826  < 2e-16 ***
## df$city2culver city -6305084     739640  -8.525 1.43e-15 ***
## df$city2palms       -6105272     786930  -7.758 2.14e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4775000 on 252 degrees of freedom
## Multiple R-squared:  0.2946, Adjusted R-squared:  0.289
## F-statistic: 52.61 on 2 and 252 DF,  p-value: < 2.2e-16
```

According to the model the city where the real state is located is significant. Beverly hills has the highest cost followed by Culver city and then Palms.

Problem 2

```
df2 <- read.csv("~/GitHub/hw1.csv")
```

a)

model1 $f(x) = b_0 + b_1 x$:

```
m1 <- lm(y~x, df2)
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 479453  479453  38.488 0.0004436 ***
## Residuals  7  87201   12457
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

model2 $f(x) = b_0 + b_1 x + b_2 x^2$:

```
m2 <- lm(y~x+I(x^2), df2)
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 479453  479453 42.0736 0.0006383 ***
## I(x^2)     1  18827   18827  1.6521 0.2460502
## Residuals  6  68374   11396
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

model3 $f(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3$:

```r
m3 <- lm(y~x+I(x^2)+I(x^3), df2)
anova(m3)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x          1 479453  479453 39.0022 0.001542 **
## I(x^2)     1  18827   18827  1.5315 0.270827
## I(x^3)     1   6909    6909  0.5620 0.487209
## Residuals  5  61465   12293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

model4 $f(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4$:

```r
m4 <- lm(y~x+I(x^2)+I(x^3)+I(x^4), df2)
anova(m4)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## x          1 479453  479453 104.7432 0.0005137 ***
## I(x^2)     1  18827   18827   4.1130 0.1124611
## I(x^3)     1   6909    6909   1.5093 0.2865864
## I(x^4)     1  43155   43155   9.4278 0.0372756 *
## Residuals  4  18310    4577
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

model5 $f(x) = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4 + b_5 x^5$:

```r
m5 <- lm(y~x+I(x^2)+I(x^3)+I(x^4)+I(x^5), df2)
anova(m5)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x          1 479453  479453 78.6485 0.003023 **
## I(x^2)     1  18827   18827  3.0883 0.177105
## I(x^3)     1   6909    6909  1.1333 0.365161
## I(x^4)     1  43155   43155  7.0791 0.076296 .
## I(x^5)     1     21      21  0.0035 0.956670
## Residuals  3  18288    6096
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b)

Based on the MSE for the training data I would choose model 4 because it has the lowest MSE.

c)

Creating testing data.

```
set.seed(456)
x=seq(0,4,by=.5)
y=500+200*x + rnorm(length(x),0,100)
df3 <- data.frame(x,y)
```

```
X_test <-seq(0,4,by=.5)
df_test <- data.frame(X_test)
```

```
myMSE <- function(arg1, arg2, n){
 s_ = (arg1-arg2)^2
 return (sum(s_)/n)
}
```

Test MSE model1:

```
m1.predictions <- predict(m1, df_test)
myMSE(df3$y, m1.predictions, nrow(df3) )
```

```
## [1] 10991.1
```

Test MSE model2:

```
m2.predictions <- predict(m2, df_test)
myMSE(df3$y, m2.predictions, nrow(df3) )
```

```
## [1] 14714.35
```

Test MSE model3:

```
m3.predictions <- predict(m3, df_test)
myMSE(df3$y, m3.predictions, nrow(df3) )
```

```
## [1] 17088.13
```

Test MSE model4:

```
m4.predictions <- predict(m4, df_test)
myMSE(df3$y, m4.predictions, nrow(df3) )
```

```
## [1] 14897.54
```

Test MSE model5:

```
m5.predictions <- predict(m5, df_test)
myMSE(df3$y, m5.predictions, nrow(df3) )
```

```
## [1] 15006.96
```

d)

The MSE for training begins to fluccuate towards smaller values as the model has higher polynomial degrees. The test MSE is optimal with the simplest model and begins to fluccuate towards larger values as the number of degrees in the polynomial increases. The MSEs make sense the data come from a linear sample. In the training case the MSE is being reduced because the model is overfitting and explaning random error. This model will fail with testing data however the simple one degree polynomial model does well in test.

Problem 3

a)

This is a regression problem and we are more interested in inference. $n = 500$,$p = 3$.

b)

This is a classification problem and we are interested in prediction. $n = 20$, $p = 13$

c)

This is a regression problem and we are interested in prediction. $n = 52$, $p = 3$

Problem 4

Given the following model $y_i = X\beta + \epsilon$ the assumptions are $E(\epsilon|X) = 0 \quad \forall X$ and $Var(\epsilon|X) = \sigma_\epsilon^2$.

If there is some lab work done and each sample contaminates another then the errors are not independent and so the variablility will not equal σ