February 14, 2018

**Assignment 1**

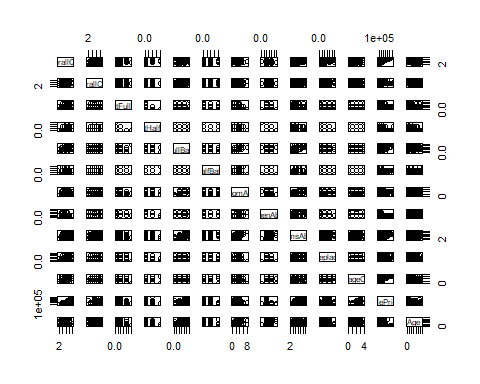
Sanchit Singhal

INF 397 – Statistical Analysis and Learning w/ Prof. Varun Rai

Spring 2018

The University of Texas at Austin

### Problem 2 - MLR on dataset  
  
# Read CSV from working directory into R  
  
MyData <- read.csv(file="austin\_house\_price.csv", header=TRUE, sep=",")  
  
# a. Scatterplot matrix with all variables in dataset  
  
pairs(MyData)



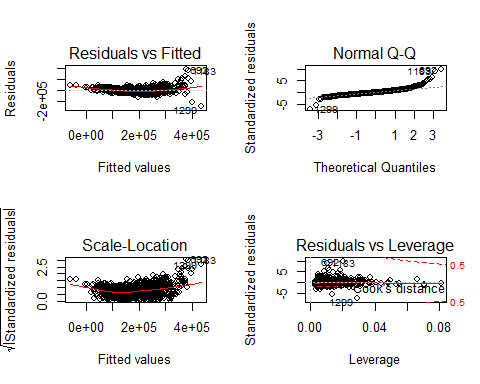
# b. Matrix of correlations of all variables  
  
cor(MyData)

## OverallQual OverallCond BsmtFullBath BsmtHalfBath FullBath  
## OverallQual 1.00000000 -0.09193234 0.11109779 -0.04015016 0.55059971  
## OverallCond -0.09193234 1.00000000 -0.05494152 0.11782092 -0.19414949  
## BsmtFullBath 0.11109779 -0.05494152 1.00000000 -0.14787096 -0.06451205  
## BsmtHalfBath -0.04015016 0.11782092 -0.14787096 1.00000000 -0.05453581  
## FullBath 0.55059971 -0.19414949 -0.06451205 -0.05453581 1.00000000  
## HalfBath 0.27345810 -0.06076933 -0.03090496 -0.01233990 0.13638059  
## BedroomAbvGr 0.10167636 0.01298006 -0.15067281 0.04651885 0.36325198  
## KitchenAbvGr -0.18388223 -0.08700086 -0.04150255 -0.03794435 0.13311521  
## TotRmsAbvGrd 0.42745234 -0.05758317 -0.05327524 -0.02383634 0.55478425  
## Fireplaces 0.39676504 -0.02381998 0.13792771 0.02897559 0.24367050  
## GarageCars 0.60067072 -0.18575751 0.13188122 -0.02089106 0.46967204  
## SalePrice 0.79098160 -0.07785589 0.22712223 -0.01684415 0.56066376  
## Age -0.57262947 0.37732550 -0.18436183 0.03605963 -0.46840292  
## HalfBath BedroomAbvGr KitchenAbvGr TotRmsAbvGrd  
## OverallQual 0.27345810 0.10167636 -0.18388223 0.42745234  
## OverallCond -0.06076933 0.01298006 -0.08700086 -0.05758317  
## BsmtFullBath -0.03090496 -0.15067281 -0.04150255 -0.05327524  
## BsmtHalfBath -0.01233990 0.04651885 -0.03794435 -0.02383634  
## FullBath 0.13638059 0.36325198 0.13311521 0.55478425  
## HalfBath 1.00000000 0.22665148 -0.06826255 0.34341486  
## BedroomAbvGr 0.22665148 1.00000000 0.19859676 0.67661994  
## KitchenAbvGr -0.06826255 0.19859676 1.00000000 0.25604541  
## TotRmsAbvGrd 0.34341486 0.67661994 0.25604541 1.00000000  
## Fireplaces 0.20364851 0.10756968 -0.12393624 0.32611448  
## GarageCars 0.21917815 0.08610644 -0.05063389 0.36228857  
## SalePrice 0.28410768 0.16821315 -0.13590737 0.53372316  
## Age -0.24272773 0.06895972 0.17591841 -0.09695522  
## Fireplaces GarageCars SalePrice Age  
## OverallQual 0.39676504 0.60067072 0.79098160 -0.57262947  
## OverallCond -0.02381998 -0.18575751 -0.07785589 0.37732550  
## BsmtFullBath 0.13792771 0.13188122 0.22712223 -0.18436183  
## BsmtHalfBath 0.02897559 -0.02089106 -0.01684415 0.03605963  
## FullBath 0.24367050 0.46967204 0.56066376 -0.46840292  
## HalfBath 0.20364851 0.21917815 0.28410768 -0.24272773  
## BedroomAbvGr 0.10756968 0.08610644 0.16821315 0.06895972  
## KitchenAbvGr -0.12393624 -0.05063389 -0.13590737 0.17591841  
## TotRmsAbvGrd 0.32611448 0.36228857 0.53372316 -0.09695522  
## Fireplaces 1.00000000 0.30078877 0.46692884 -0.14854356  
## GarageCars 0.30078877 1.00000000 0.64040920 -0.53872739  
## SalePrice 0.46692884 0.64040920 1.00000000 -0.52335042  
## Age -0.14854356 -0.53872739 -0.52335042 1.00000000

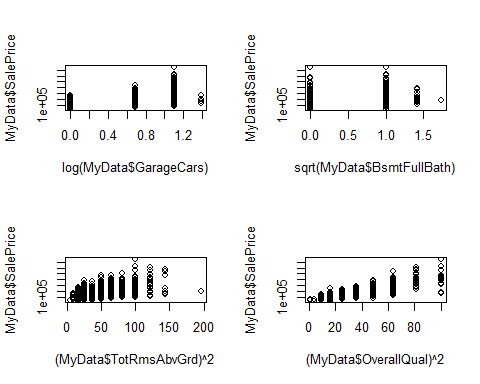
# c. Multiple Linear Regression  
  
lm.fit=lm(SalePrice~., data=MyData)  
summary(lm.fit)

##   
## Call:  
## lm(formula = SalePrice ~ ., data = MyData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -274626 -21629 -3288 17476 374855   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -83029.36 10280.94 -8.076 1.40e-15 \*\*\*  
## OverallQual 23140.58 1197.68 19.321 < 2e-16 \*\*\*  
## OverallCond 4340.82 1035.88 4.190 2.95e-05 \*\*\*  
## BsmtFullBath 21740.63 2130.05 10.207 < 2e-16 \*\*\*  
## BsmtHalfBath 10236.97 4429.58 2.311 0.021 \*   
## FullBath 13417.14 2825.20 4.749 2.25e-06 \*\*\*  
## HalfBath 239.56 2329.34 0.103 0.918   
## BedroomAbvGr -9599.12 1841.24 -5.213 2.12e-07 \*\*\*  
## KitchenAbvGr -30303.01 5344.86 -5.670 1.73e-08 \*\*\*  
## TotRmsAbvGrd 15129.23 1159.76 13.045 < 2e-16 \*\*\*  
## Fireplaces 12668.63 1836.60 6.898 7.87e-12 \*\*\*  
## GarageCars 16766.94 1873.62 8.949 < 2e-16 \*\*\*  
## Age -248.83 53.59 -4.643 3.75e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 39330 on 1447 degrees of freedom  
## Multiple R-squared: 0.7569, Adjusted R-squared: 0.7549   
## F-statistic: 375.4 on 12 and 1447 DF, p-value: < 2.2e-16

par(mfrow = c(2, 2))  
plot(lm.fit)



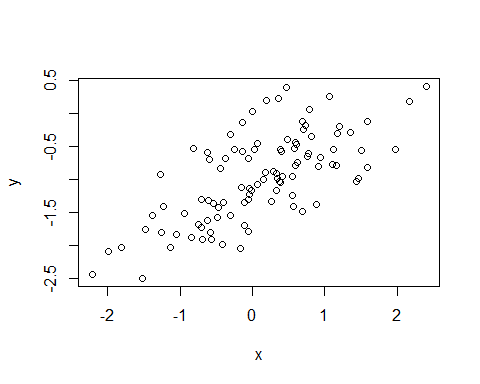
# Relationship between predictors and response:   
  
 # By testing the null hypothesis of that there is no relationship, we can  
 # reject it by looking at the p-value corresponding to the F-statistic. In  
 # this case, it is very small (<2.2e-16) which means there appears to be a  
 # strong relationship between "SalePrice" and atleast some of the predictors.  
 # Indeed, by looking at the regression coefficients it can be seen that  
 # "GarageCars", "BsmtFullBath", "TotRmsAbvGrd", "OverallQual" all have small  
 # p-values and are therefore statistically significant.  
  
 # Coefficient for the age variable:  
   
 # The regression coefficient for the age, -248.83, suggests that for every 1  
 # unit in age (presumably a year), SalePrice decreases by the coefficient. In  
 # other words, the price falls every year which makes sense because property is  
 # usuallly more expensive the newer it is.  
  
  
# d. Transformation of the variables  
  
par(mfrow = c(2, 2))  
plot(log(MyData$GarageCars), MyData$SalePrice)  
plot(sqrt(MyData$BsmtFullBath), MyData$SalePrice)  
plot((MyData$TotRmsAbvGrd)^2, MyData$SalePrice)  
plot((MyData$OverallQual)^2, MyData$SalePrice)



# Comment on findings:  
  
 # I decided to transform variables that had the highest statistically  
 # significance (lowest p-values) because they have the greastest impact on the  
 # SalesPrice. After trying out some transformation, I believe the square of the  
 # overall quality gives the most linear looking plot.  
  
  
### Problem 3 - SLR on simulated data  
  
set.seed(1)  
par(mfrow = c(1, 1))  
  
# a. Generation of Feature X  
  
x = rnorm(100)  
  
# b. Generation of Feature eps  
  
eps = rnorm(100, 0, sqrt(0.25))  
  
# c. Generation of response  
  
y = y = -1 + 0.5\*x + eps  
length(y)

## [1] 100

# Length of vector, Y:  
   
 # The length of vector, Y, is 100 which makes sense since it a linear function  
 # of 2 sets of 100 values  
  
 # Values for B0 & B1:  
  
 # B0 = -1, B1 = 0.5 as seen from the original equation  
  
  
# d. Scatterplot  
  
plot(x, y)



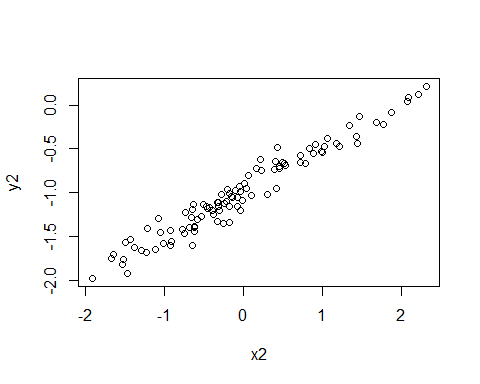
# Comment on observations:  
   
 # The relationship between x & y has a positive, linear slope with some  
 # variance due to the noise introduced by the eps variable.  
  
# e. Least Square Linear Model  
  
lm.fit2 <- lm(y ~ x)  
summary(lm.fit2)

##   
## Call:  
## lm(formula = y ~ x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.93842 -0.30688 -0.06975 0.26970 1.17309   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.01885 0.04849 -21.010 < 2e-16 \*\*\*  
## x 0.49947 0.05386 9.273 4.58e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4814 on 98 degrees of freedom  
## Multiple R-squared: 0.4674, Adjusted R-squared: 0.4619   
## F-statistic: 85.99 on 1 and 98 DF, p-value: 4.583e-15

# Comment on Model:  
   
 # The model has a large F-statistic with a small p-value (4.583e-15) and so the null  
 # hypothesis can be rejected. This makes sense to me as we know y was indeed  
 # generated using x and therefore, the two definitively have a relationship.  
  
 # How do B^0 and B^1 compare to B0 and B1:  
   
 # The constructed values for B^0 (-1.019) and B^1 (0.499) were very close to  
 # the true values of -1 and 0.5. This means the linear regression model does a  
 # great job modelling the relationship between x & y.  
  
# f. Polynomial Regression Model  
  
lm.fit2\_sq = lm(y~x+I(x^2))  
summary(lm.fit2\_sq)

##   
## Call:  
## lm(formula = y ~ x + I(x^2))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.98252 -0.31270 -0.06441 0.29014 1.13500   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.97164 0.05883 -16.517 < 2e-16 \*\*\*  
## x 0.50858 0.05399 9.420 2.4e-15 \*\*\*  
## I(x^2) -0.05946 0.04238 -1.403 0.164   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.479 on 97 degrees of freedom  
## Multiple R-squared: 0.4779, Adjusted R-squared: 0.4672   
## F-statistic: 44.4 on 2 and 97 DF, p-value: 2.038e-14

# Does quadratic term improve the model fit:  
   
 # There is evidence that the model fit has increased slightly as the RSE has  
 # decreased and the R^2 is higher. However, when taking into account the large  
 # p-value for the x^2 coefficient, it can be concluded that x^2 does not have  
 # a relationship with y and the model is most likely overfitting the training  
 # data by learning too much of the noise.  
  
# g. Reduction of Noise  
  
set.seed(1)  
eps2 = rnorm(100, 0, 0.125)  
x2 = rnorm(100)  
y2 = -1 + 0.5\*x2 + eps2  
plot(x2, y2)



lm.fit3 = lm(y2~x2)  
summary(lm.fit3)

##   
## Call:  
## lm(formula = y2 ~ x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.29052 -0.07545 0.00067 0.07288 0.28664   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.98639 0.01129 -87.34 <2e-16 \*\*\*  
## x2 0.49988 0.01184 42.22 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1128 on 98 degrees of freedom  
## Multiple R-squared: 0.9479, Adjusted R-squared: 0.9474   
## F-statistic: 1782 on 1 and 98 DF, p-value: < 2.2e-16

# Description of Results  
   
 # By decreasing the variance of the normal distribution that generates the  
 # error term, eps, we are able to reduce noise. The coefficients for B0 and B1  
 # remain very similar which tells us that the model remained the same.  
 # However, the RSE has significantly decreased, and R^2 has increased which  
 # means the model fits extremely well. Again, this makes sense because the  
 # underlying data is near-perfect with very little error.