Lab No. 7

Elizabeth Walter

```
knitr::opts_chunk$set(echo = TRUE,
                      fig.align = 'center')
```

```
library(tidyverse)
library(ggplot2)
ameslist <- read.csv("C:/Users/walte/Desktop/MSU SSQDA/SSC 442/Data/ames.csv",
                 header = TRUE,
                 sep = ",")
b <- which(sapply(ameslist, class) %in% c('integer'))
Ames <- ameslist[names(ameslist[b])]
```

1. Use the lm() function in a simple linear regression (e.g., with only one predictor) with SalePrice as the response to determine the value of a garage.

Our simple regression suggests that the average sale price for a house with no garage is 71,357.42 USD and a 1 sq ft increase in garage space is correlated with an expected increase in the sale price of the house by 231.65 USD, and the effect is statistically significant.

```
lm.fit = lm(SalePrice ~ GarageArea, data = Ames)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = SalePrice ~ GarageArea, data = Ames)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -279451  -33024   -5045   24479  490913
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71357.421   3949.003   18.07   <2e-16 ***
## GarageArea    231.646      7.608   30.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62140 on 1458 degrees of freedom
## Multiple R-squared:  0.3887, Adjusted R-squared:  0.3882
## F-statistic:   927 on 1 and 1458 DF,  p-value: < 2.2e-16
```

2. Use the lm() function to perform a multiple linear regression with SalePrice as the response and all other variables from your Ames data as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

a. Is there a relationship between the predictors and the response?

Many of the predictors with the most significance are all positive, but there are also statistically significant predictors that have a negative relationship to SalePrice. There is great range in magnitude of the estimated effect of the predictors on SalePrice, ranging from 1's - 10000's.

b. Which predictors appear to have a statistically significant relationship to the response?

LotArea, OverallQual, OveralCond, Year Built, MasVnrArea, BsmtFinSF1, X1stFlrSF, X2ndFlrSF, BsmtFullBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageCars, WoodDeckSF, ScreenPorch, PoolArea.

c. What does the coefficient for the year variable suggest?

The coefficient for year sold is -253.6, suggesting that all else constant, the sale price of a house decreased by $253.6, on average, each year beyond 2006 (until the year it was sold). However we see it is not statistically significant.

```
Ames$MSSubClass <- as.numeric(Ames$MSSubClass)
mult_lm <- lm(SalePrice ~ MSSubClass + LotFrontage + LotArea + OverallQual + OverallCond + YearBuilt + Y

summary(mult_lm)
```

```
##
## Call:
## lm(formula = SalePrice ~ MSSubClass + LotFrontage + LotArea +
##     OverallQual + OverallCond + YearBuilt + YearRemodAdd + MasVnrArea +
##     BsmtFinSF1 + BsmtFinSF2 + BsmtUnfSF + TotalBsmtSF + X1stFlrSF +
##     X2ndFlrSF + LowQualFinSF + GrLivArea + BsmtFullBath + BsmtHalfBath +
##     FullBath + HalfBath + BedroomAbvGr + KitchenAbvGr + TotRmsAbvGrd +
##     Fireplaces + GarageYrBlt + GarageCars + GarageArea + WoodDeckSF +
##     OpenPorchSF + EnclosedPorch + X3SsnPorch + ScreenPorch +
##     PoolArea + MiscVal + MoSold + YrSold, data = Ames)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -442865  -16873   -2581   14998  318042
##
## Coefficients: (2 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -3.232e+05  1.701e+06  -0.190 0.849317
## MSSubClass    -2.005e+02  3.449e+01  -5.814 8.03e-09 ***
## LotFrontage   -1.161e+02  6.124e+01  -1.896 0.058203 .
## LotArea        5.454e-01  1.573e-01   3.466 0.000548 ***
## OverallQual    1.870e+04  1.478e+03  12.646  < 2e-16 ***
## OverallCond    5.227e+03  1.367e+03   3.824 0.000139 ***
## YearBuilt      3.170e+02  8.762e+01   3.617 0.000311 ***
## YearRemodAdd   1.206e+02  8.661e+01   1.392 0.164174    
## MasVnrArea     3.160e+01  7.006e+00   4.511 7.15e-06 ***
## BsmtFinSF1     1.739e+01  5.835e+00   2.980 0.002947 ** 
## BsmtFinSF2     8.362e+00  8.763e+00   0.954 0.340205    
## BsmtUnfSF      5.006e+00  5.275e+00   0.949 0.342890    
## TotalBsmtSF           NA         NA      NA       NA    
## X1stFlrSF      4.591e+01  7.356e+00   6.241 6.21e-10 ***
## X2ndFlrSF      4.668e+01  6.099e+00   7.654 4.28e-14 ***
## LowQualFinSF   3.415e+01  2.788e+01   1.225 0.220788    
## GrLivArea             NA         NA      NA       NA    
## BsmtFullBath   8.980e+03  3.194e+03   2.812 0.005018 ** 
## BsmtHalfBath   2.490e+03  5.071e+03   0.491 0.623487    
## FullBath       5.390e+03  3.529e+03   1.527 0.126941    
## HalfBath      -1.119e+03  3.320e+03  -0.337 0.736244    
## BedroomAbvGr  -1.023e+04  2.154e+03  -4.750 2.30e-06 ***
## KitchenAbvGr  -2.193e+04  6.704e+03  -3.271 0.001105 ** 
## TotRmsAbvGrd   5.440e+03  1.486e+03   3.661 0.000263 ***
## Fireplaces     4.375e+03  2.188e+03   2.000 0.045793 *  
## GarageYrBlt   -4.914e+01  9.093e+01  -0.540 0.589011    
## GarageCars     1.679e+04  3.487e+03   4.815 1.68e-06 ***
## GarageArea     6.488e+00  1.211e+01   0.536 0.592338    
## WoodDeckSF     2.155e+01  1.002e+01   2.151 0.031713 *  
## OpenPorchSF   -2.315e+00  1.948e+01  -0.119 0.905404    
## EnclosedPorch  7.233e+00  2.061e+01   0.351 0.725733    
## X3SsnPorch     3.458e+01  3.749e+01   0.922 0.356593    
## ScreenPorch    5.797e+01  2.040e+01   2.842 0.004572 ** 
## PoolArea      -6.126e+01  2.984e+01  -2.053 0.040326 *  
## MiscVal       -3.850e+00  6.955e+00  -0.554 0.579980    
## MoSold        -2.240e+02  4.227e+02  -0.530 0.596213    
## YrSold        -2.536e+02  8.454e+02  -0.300 0.764216    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 36790 on 1086 degrees of freedom
##   (339 observations deleted due to missingness)
## Multiple R-squared:  0.8095, Adjusted R-squared:  0.8036 
## F-statistic: 135.7 on 34 and 1086 DF,  p-value: < 2.2e-16
```

3. Use the : symbols to fit a linear regression model with one well-chosen interaction effects. Why did you do this?

I wanted to use two terms that appeared to be statistically significant on their own. I noticed the NA reported in the summary() for TotalBsmtSF and GrLivArea due to high correlation between the variables, which makes sense- If a house has a basement, the floor plan is often the same or nearly the same for it as the floor plan for the first floor or more, so it is likely that the variable that contains the combined area of the first and all higher floors, GrLivArea, will contain TotalBsmtSF in that combination. Including the interaction effect in a model of TotalBsmtSF and GrLivArea on SalePrice revealed that the estimated change in average sale price for an increase of above ground area decreased for larger total basement areas, and vice versa. This makes sense that more area on either floor - and therefore greater total area of the house - would decrease the value of an additional unit on the other. For that reason, I wanted to test again for something slightly more nuanced.

```
sp_bm_liv <- lm(SalePrice ~ TotalBsmtSF + GrLivArea + TotalBsmtSF:GrLivArea, data = Ames)
summary(sp_bm_liv)
```

```
##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF + GrLivArea + TotalBsmtSF:GrLivArea,
##     data = Ames)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -272782  -23016    -416   21742  316301
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -7.133e+04  6.552e+03  -10.89   <2e-16 ***
## TotalBsmtSF            1.170e+02  5.480e+00   21.35   <2e-16 ***
## GrLivArea              1.130e+02  3.826e+00   29.53   <2e-16 ***
## TotalBsmtSF:GrLivArea -2.497e-02  2.203e-03  -11.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47660 on 1456 degrees of freedom
## Multiple R-squared:  0.6408, Adjusted R-squared:  0.6401
## F-statistic: 865.8 on 3 and 1456 DF,  p-value: < 2.2e-16
```

```
coef(sp_bm_liv)
```

```
##           (Intercept)           TotalBsmtSF             GrLivArea
##         -7.133387e+04          1.169705e+02          1.129711e+02
## TotalBsmtSF:GrLivArea
##         -2.497334e-02
```

```
b1 <- coef(sp_bm_liv)[2]
b2 <- coef(sp_bm_liv)[3]
b3 <- coef(sp_bm_liv)[4]

ch_x1 <- function(x2){
  ch_x1 <- b1 + (b3*x2)
```

```r
  return(ch_x1)
}
ch_x2 <- function(x1){
  ch_x2 <- b2 + (b3*x1)
  return(ch_x2)
}
ch_x1(500)
```

```
## TotalBsmtSF
##    104.4838
```

```r
ch_x1(750)
```

```
## TotalBsmtSF
##    98.24049
```

```r
ch_x1(1000)
```

```
## TotalBsmtSF
##    91.99716
```

```r
ch_x2(500)
```

```
## GrLivArea
##  100.4845
```

```r
ch_x2(750)
```

```
## GrLivArea
##  94.24112
```

```r
ch_x2(1000)
```

```
## GrLivArea
##  87.99779
```

Thinking that there is some interaction between the quality of a house and the age of the house, I decided to look at the estimated interaction effect of OverallQual & YearBuilt. Here it was interesting to see the coefficient estimates of both OverallQual and YearBuilt change sign in the simple model interaction as compared to the big model from question 2, which is likely due to the simpler model attributing negative effects of other predictors on Sale Price to OverallQual and YearBuilt. However we see that the coefficient estimate of the interaction term is positive. This indicates that, in terms of our estimates, for a one unit increase in quality score, the change in the average sale price is larger for newer houses, and for a one year increase in build year, the change in the average sale price is larger for higher quality houses.

```r
sp_oq_yb <- lm(SalePrice ~ OverallQual + YearBuilt + OverallQual:YearBuilt, data = Ames)
summary(sp_oq_yb)
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + YearBuilt + OverallQual:YearBuilt,
##     data = Ames)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -223855  -27787   -3031   18388  389298
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2563675.77  381274.65   6.724 2.53e-11 ***
## OverallQual           -478341.87   60454.89  -7.912 4.96e-15 ***
## YearBuilt               -1340.13     194.04  -6.907 7.40e-12 ***
## OverallQual:YearBuilt     263.67      30.63   8.609  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47000 on 1456 degrees of freedom
## Multiple R-squared:  0.6508, Adjusted R-squared:   0.65
## F-statistic: 904.3 on 3 and 1456 DF,  p-value: < 2.2e-16
```

```
coef(sp_oq_yb)
```

```
##           (Intercept)           OverallQual             YearBuilt
##          2563675.7723          -478341.8671            -1340.1264
## OverallQual:YearBuilt
##              263.6739
```

```
b1 <- coef(sp_oq_yb)[2]
b2 <- coef(sp_oq_yb)[3]
b3 <- coef(sp_oq_yb)[4]

ch_x1(1960)
```

```
## OverallQual
##    38458.94
```

```
ch_x1(1970)
```

```
## OverallQual
##    41095.68
```

```
ch_x1(1980)
```

```
## OverallQual
##    43732.42
```

```
ch_x2(3)
```

```
## YearBuilt
## -549.1048
```

```
ch_x2(5)
```

```
## YearBuilt
## -21.75704
```

```
ch_x2(7)
```

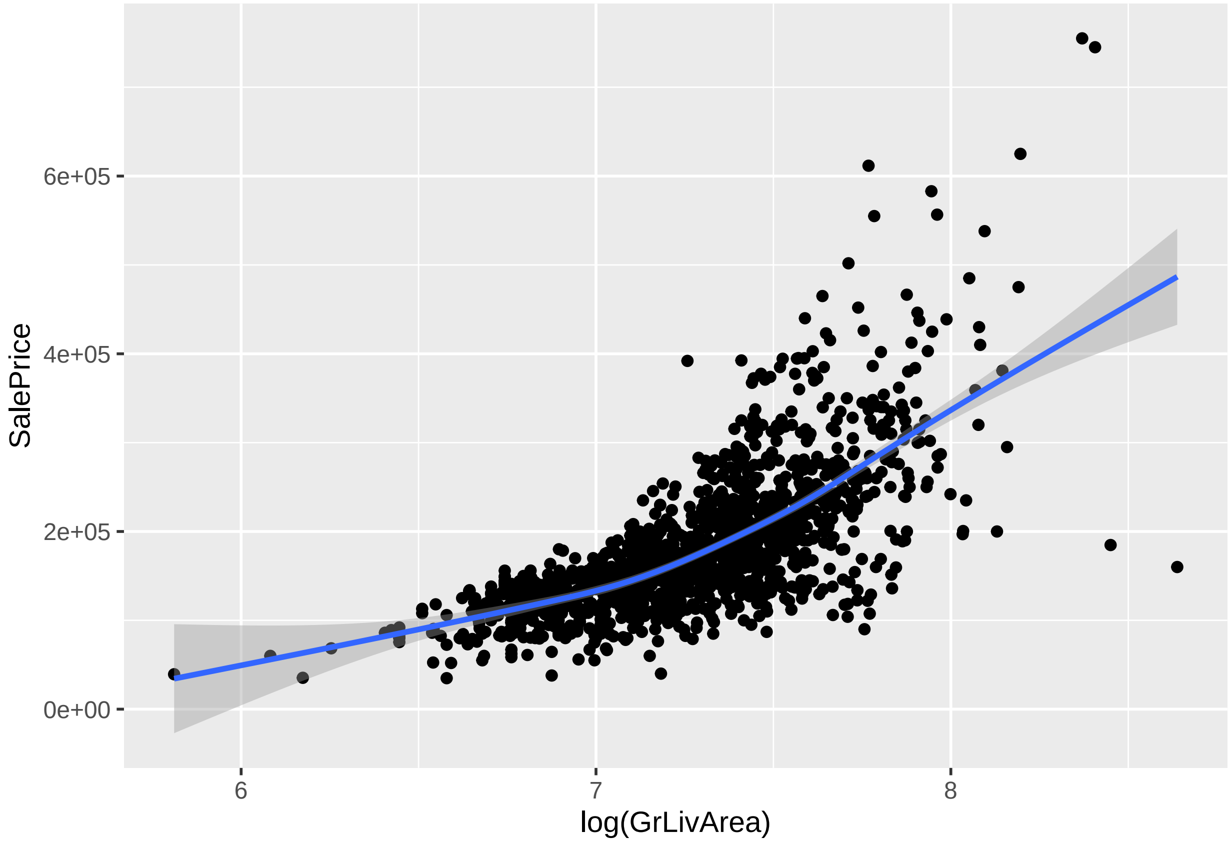
```
## YearBuilt
##  505.5907
```

4. Try two different transformations of the variables, such as ln(x), x^2, sqrt(x). Do any of these make sense to include in a model of SalePrice? Comment on your findings.

With this data, I do not see how a transformation of sqrt() or ^2 on any of the variables can be helpful. However, I believe that the ln() transformation could be helpful for reducing the impact of very large/small outliers when trying to analyze a regression line.

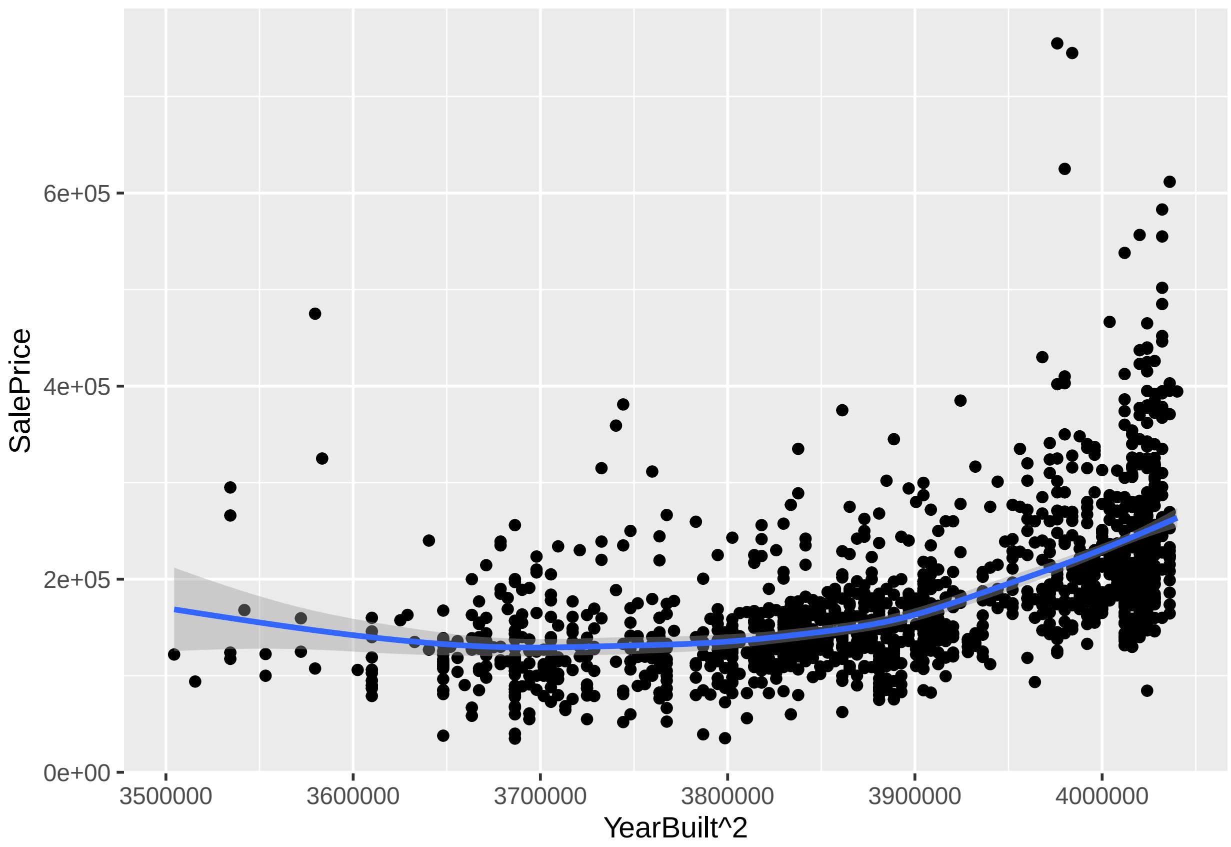
```
ggplot(Ames, aes(x = log(GrLivArea), y = SalePrice)) +
  geom_point() + geom_smooth()
```

```
## 'geom_smooth()' using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



```
ggplot(Ames, aes(x = YearBuilt**2, y = SalePrice)) +
  geom_point() + geom_smooth()
```

```
## 'geom_smooth()' using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```


SalePrice
6e+05
4e+05
2e+05
0e+00
3500000
3600000
3700000
3800000
3900000
4000000
YearBuilt^2