Final Exam - Take Home

Grant Herrenbruck

2022-12-10

Problem 1

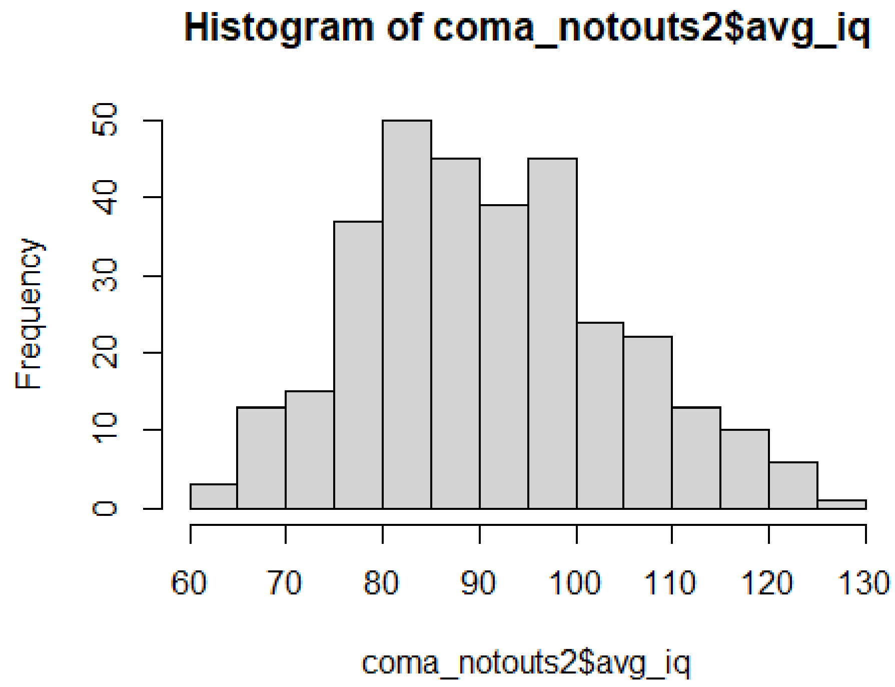
```
# Remove identifier column
coma <- read.csv("Wong.csv")[,-1]

# Combine the two IQ scores to get an average
coma$avg_iq <- (coma$piq + coma$viq)/2

# log average iq score
coma$log.avg_iq <- log(coma$avg_iq, 10)

# remove outlier
coma_notouts <- coma[-331,]

# remove duration outliers
coma_notouts2 <- filter(coma_notouts, duration < 100)
```

Part A

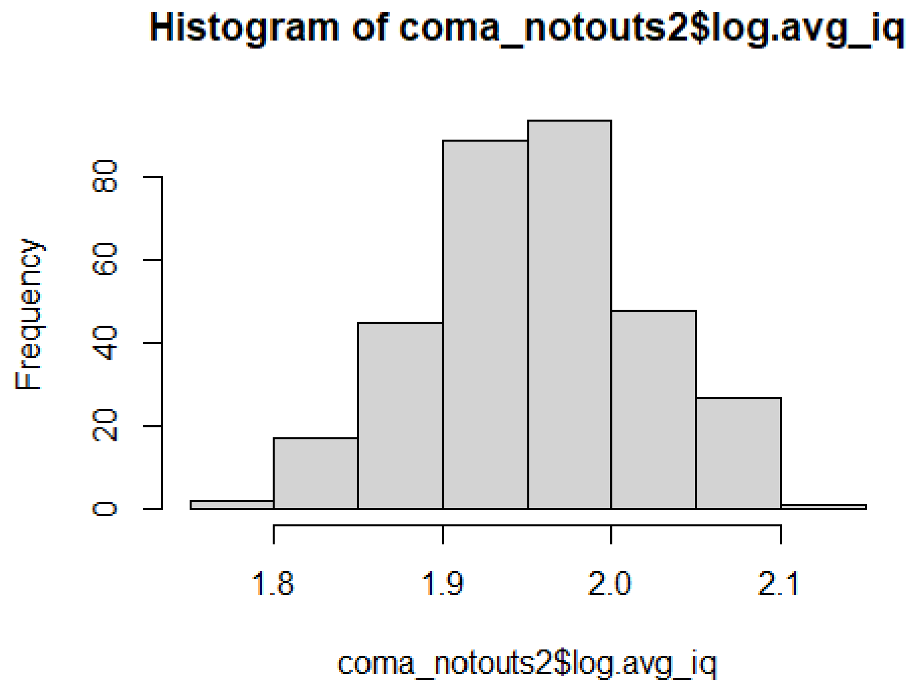
Step 1 - Create average IQ, remove identifier column For my response variable, I chose to analyze the average between the math iq score and the verbal iq score to try and get an all-encompassing single predictor. Along with this, I removed the ID column from the data set since it has no predictive importance on IQ score.

Step 2 - Take the log of average IQ score Even though the normal average IQ scores maintained a fairly normal distribution, I chose to take a log transformation to enhance the predictive power of the confidence intervals in later analysis.

```
hist(coma_notouts2$avg_iq)
```



```
hist(coma_notouts2$log.avg_iq)
```

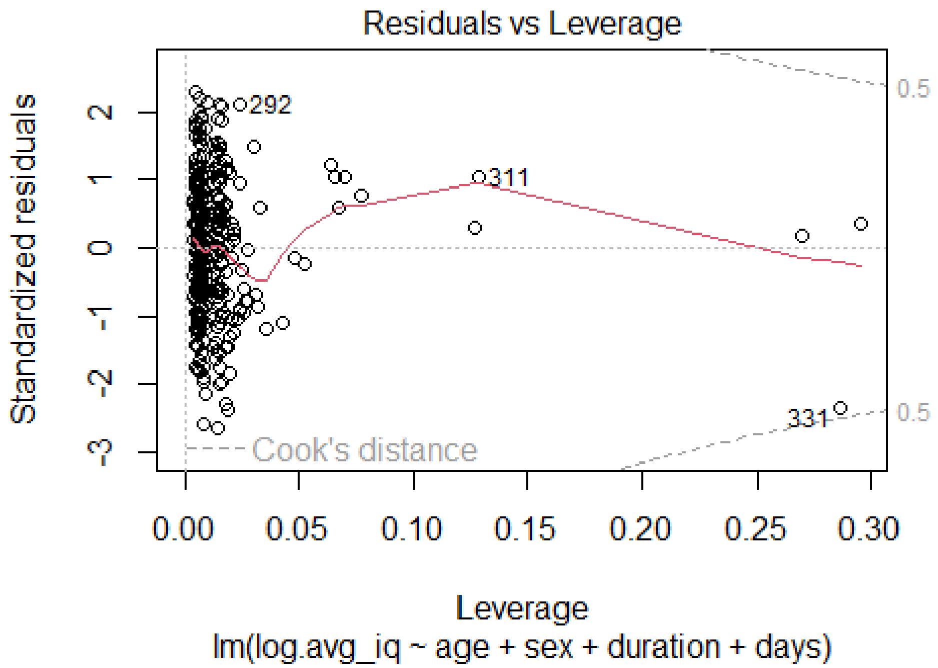


Step 3 - Fit full log model, run diagnostics, remove outlier

Based on the Cook's Distance plot, we see that observation 331 was highly leveraged and possibly influencing the output. After removing this observation, the days post coma and duration of coma's predictive power on log average IQ became significant, proving that this observation was indeed affecting the output.

```
j2 <- lm(log.avg_iq ~ age+sex+duration+days, coma)
summary(j2)

## 
## Call:
## lm(formula = log.avg_iq ~ age + sex + duration + days, data = coma)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.164438 -0.040486 -0.001289  0.044057  0.143119 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.951e+00  1.116e-02 174.747   <2e-16 ***
## age          2.482e-04  2.552e-04   0.972    0.332    
## sexMale     -1.228e-03  8.501e-03  -0.144    0.885    
## duration    -2.268e-04  1.384e-04  -1.638    0.102    
## days         2.718e-06  3.146e-06   0.864    0.388    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.06295 on 326 degrees of freedom
## Multiple R-squared:  0.01329,    Adjusted R-squared:  0.001182 
## F-statistic: 1.098 on 4 and 326 DF,  p-value: 0.3577
```

```
j3 <- lm(log.avg_iq ~ age+sex+duration+days, coma_notouts)
summary(j3)

## 
## Call:
## lm(formula = log.avg_iq ~ age + sex + duration + days, data = 
coma_notouts)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.163949 -0.040511 -0.001193  0.044146  0.142557 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.950e+00  1.109e-02 175.951   <2e-16 ***
## age          2.268e-04  2.536e-04   0.895   0.3717    
## sexMale     -1.125e-03  8.442e-03  -0.133   0.8940    
## duration    -2.842e-04  1.396e-04  -2.036   0.0426 *  
## days         7.314e-06  3.675e-06   1.990   0.0474 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0625 on 325 degrees of freedom
## Multiple R-squared:  0.02284,    Adjusted R-squared:  0.01082
## F-statistic: 1.899 on 4 and 325 DF,  p-value: 0.1103

plot(j2, which = 5)
```



Step 4 - Test for any interaction

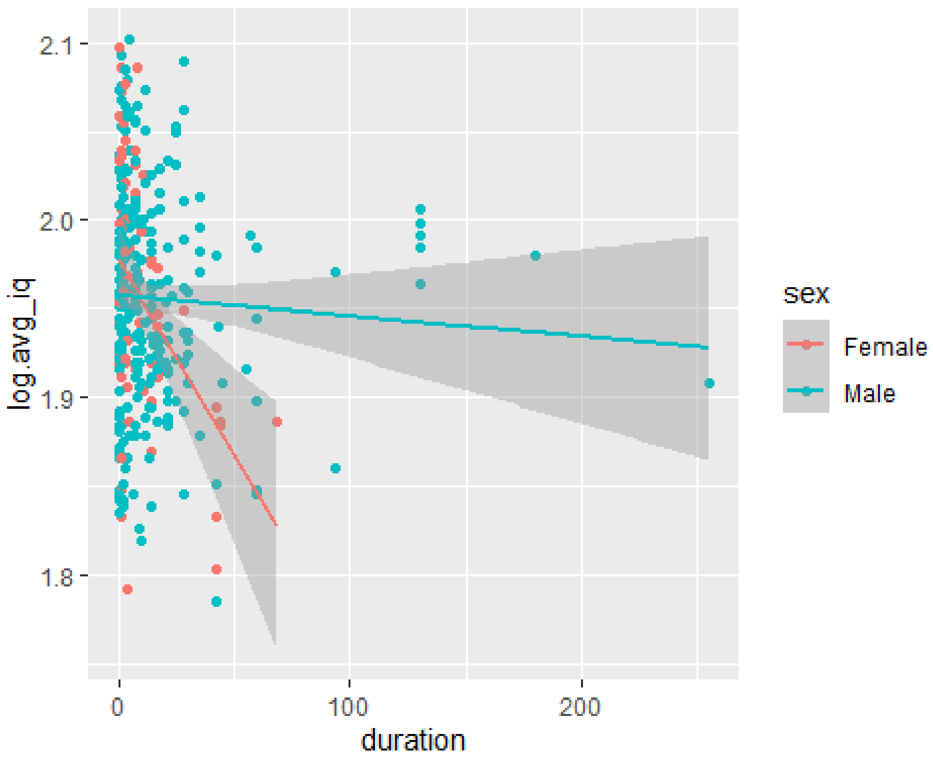
After testing for any significant interactions, I was able to find a significant interaction effect between the duration of a given coma and their gender. By implementing this interaction into the model, the "genderMale" predictor became significant, telling us that if the subject who was in a coma was a male, their estimated log average IQ would be .0021 lower than if the subject were a female, holding other variables constant. In addition, the interaction effect is telling us that the effect the duration of being in a coma has on an individual's estimated log average IQ score depends on if they are male or female.

```
j4 <- lm(log.avg_iq ~ age+sex+duration+days+duration*sex, coma_notouts)
summary(j4)

##
## Call:
## lm(formula = log.avg_iq ~ age + sex + duration + days + duration *
##     sex, data = coma_notouts)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.174423 -0.039396  0.000191  0.043743  0.143929 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       1.970e+00  1.209e-02 162.969  < 2e-16 ***
## age               1.989e-04  2.491e-04   0.799 0.425019    
## sexMale          -2.109e-02  9.945e-03  -2.121 0.034684 *  
## duration         -2.253e-03  5.590e-04  -4.030 6.95e-05 ***
## days              7.359e-06  3.608e-06   2.039 0.042219 *  
## sexMale:duration  2.080e-03  5.726e-04   3.633 0.000326 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.06136 on 324 degrees of freedom
## Multiple R-squared:  0.06108,    Adjusted R-squared:  0.0466 
## F-statistic: 4.216 on 5 and 324 DF,  p-value: 0.0009998
```

However, after looking at the plot output of the interaction, I had suspicions that the data could be creating a false interaction between duration and gender and their effect on estimated log average IQ. The plot below shows that there are no observed women that were in a coma for more than about 75 days, but there are coma durations up to 250 days for the men observed. To validate the conclusion that the coma duration's effect on log average IQ score depends on gender, I felt that I had to remove the outlying durations to get an equal comparison of duration between genders and their effect on log average IQ.

```
qplot(x=duration,y=log.avg_iq, data=coma_notouts, color=sex) +
geom_smooth(method="lm")

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



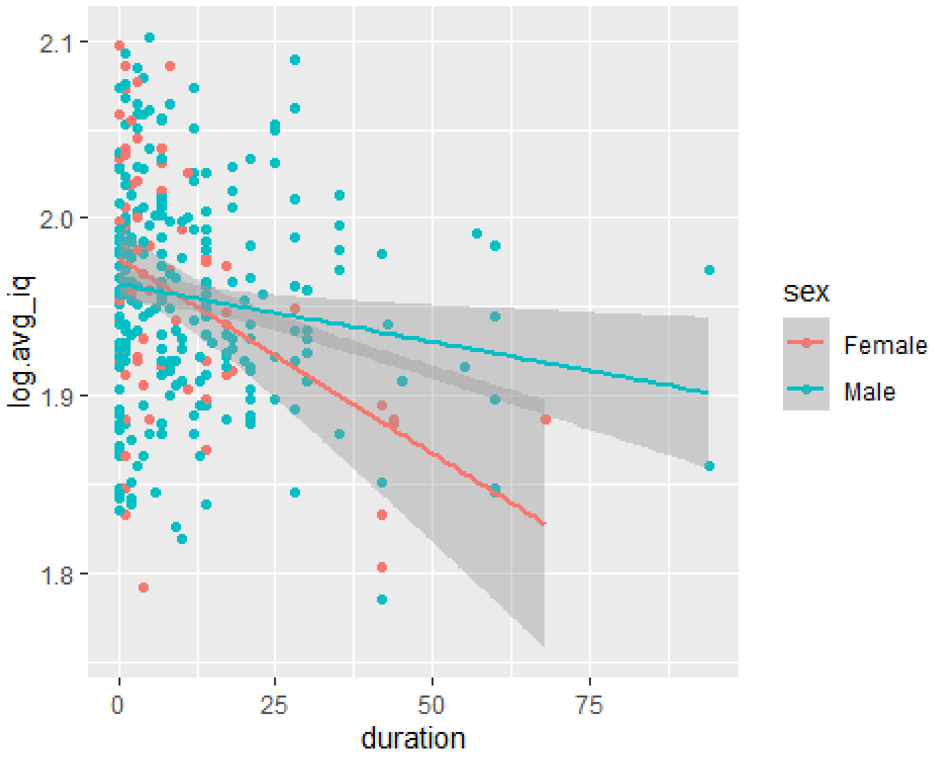
After removing the outlying values for duration, the interaction effect between duration and gender still held. This tells us that effect the duration has on the log average IQ score does in fact depend on if the subject is male or female, holding other variables constant.

An interesting output that changed after removing the duration outliers was the “genderMale” predictor. The “genderMale” predictor became insignificant, telling us that the post coma log average IQ score no longer depends on if the subject is male or female, holding other variables constant.

```
j5 <- lm(log.avg_iq ~ age+sex+duration+days+duration*sex, coma_notouts2)
summary(j5)

##
## Call:
## lm(formula = log.avg_iq ~ age + sex + duration + days + duration *
##     sex, data = coma_notouts2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17480 -0.04052 -0.00159  0.04388  0.14149
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.972e+00  1.211e-02 162.839  < 2e-16 ***
## age                  1.053e-04  2.510e-04   0.420   0.6751
## sexMale             -1.525e-02  1.020e-02  -1.494   0.1361
## duration            -2.274e-03  5.581e-04  -4.074 5.84e-05 ***
## days                 7.905e-06  3.702e-06   2.135   0.0335 *
## sexMale:duration  1.513e-03  6.134e-04   2.467   0.0142 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06126 on 317 degrees of freedom
## Multiple R-squared:  0.07802,    Adjusted R-squared:  0.06348
## F-statistic: 5.365 on 5 and 317 DF,  p-value: 9.454e-05
```

```
qplot(x=duration,y=log.avg_iq, data=coma_notouts2, color=sex) +
geom_smooth(method="lm")
```

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



Part B

Here we are using the final model that was described above to see if IQ recovers over time. For this problem, I chose to analyze 6 different time frames post coma - 50 days, 200 days, 350 days, 500 days, 650 days, and 800 days.

My results showed that the estimated median average IQ, holding other variables considered constant, does in fact recover over time by a small amount.

```
j5 <- lm(log.avg_iq ~ age+sex+duration+days+duration*sex, coma_notouts2)

pred.coma50 <- data.frame(age = coma_notouts2$age, sex = coma_notouts2$sex,
                        days = 50, duration = coma_notouts2$duration)
pred.coma200 <- data.frame(age = coma_notouts2$age, sex = coma_notouts2$sex,
                        days = 200, duration = coma_notouts2$duration)
pred.coma350 <- data.frame(age = coma_notouts2$age, sex = coma_notouts2$sex,
                        days = 350, duration = coma_notouts2$duration)
pred.coma500 <- data.frame(age = coma_notouts2$age, sex = coma_notouts2$sex,
                        days = 500, duration = coma_notouts2$duration)
pred.coma650 <- data.frame(age = coma_notouts2$age, sex = coma_notouts2$sex,
                        days = 650, duration = coma_notouts2$duration)
pred.coma800 <- data.frame(age = coma_notouts2$age, sex = coma_notouts2$sex,
                        days = 800, duration = coma_notouts2$duration)
# 50
days_50.med <- mean(10^predict(j5, pred.coma50, interval = "confidence")[,1])
days_50.lwr <- mean(10^predict(j5, pred.coma50, interval = "confidence")[,2])
days_50.upr <- mean(10^predict(j5, pred.coma50, interval = "confidence")[,3])
days_pc <- data.frame(days_50.med, days_50.lwr, days_50.upr)
colnames(days_pc) <- c("Median", "Lower", "Upper")
rownames(days_pc) <- "50 days post coma"

# 200
days_200.med <- mean(10^predict(j5, pred.coma200, interval =
"confidence")[,1])
days_200.lwr <- mean(10^predict(j5, pred.coma200, interval =
"confidence")[,2])
days_200.upr <- mean(10^predict(j5, pred.coma200, interval =
"confidence")[,3])
days_200 <- data.frame(days_200.med, days_200.lwr, days_200.upr)
colnames(days_200) <- c("Median", "Lower", "Upper")
rownames(days_200) <- "200 days post coma"

# 350
days_350.med <- mean(10^predict(j5, pred.coma350, interval =
"confidence")[,1])
days_350.lwr <- mean(10^predict(j5, pred.coma350, interval =
"confidence")[,2])
days_350.upr <- mean(10^predict(j5, pred.coma350, interval =
"confidence")[,3])
days_350 <- data.frame(days_350.med, days_350.lwr, days_350.upr)
colnames(days_350) <- c("Median", "Lower", "Upper")
rownames(days_350) <- "350 days post coma"


# 500
days_500.med <- mean(10^predict(j5, pred.coma500, interval =
"confidence")[,1])
days_500.lwr <- mean(10^predict(j5, pred.coma500, interval =
"confidence")[,2])
days_500.upr <- mean(10^predict(j5, pred.coma500, interval =
"confidence")[,3])
```

```
days_500 <- data.frame(days_500.med, days_500.lwr, days_500.upr)
colnames(days_500) <- c("Median", "Lower", "Upper")
rownames(days_500) <- "500 days post coma"


# 650
days_650.med <- mean(10^predict(j5, pred.coma650, interval =
"confidence")[,1])
days_650.lwr <- mean(10^predict(j5, pred.coma650, interval =
"confidence")[,2])
days_650.upr <- mean(10^predict(j5, pred.coma650, interval =
"confidence")[,3])
days_650 <- data.frame(days_650.med, days_650.lwr, days_650.upr)
colnames(days_650) <- c("Median", "Lower", "Upper")
rownames(days_650) <- "650 days post coma"


# 800
days_800.med <- mean(10^predict(j5, pred.coma800, interval =
"confidence")[,1])
days_800.lwr <- mean(10^predict(j5, pred.coma800, interval =
"confidence")[,2])
days_800.upr <- mean(10^predict(j5, pred.coma800, interval =
"confidence")[,3])
days_800 <- data.frame(days_800.med, days_800.lwr, days_800.upr)
colnames(days_800) <- c("Median", "Lower", "Upper")
rownames(days_800) <- "800 days post coma"

days_pc <- rbind(days_pc, days_200, days_350, days_500, days_650, days_800)
days_pc
```

```
##                       Median    Lower    Upper
## 50 days post coma  89.74337 86.95387 92.64043
## 200 days post coma 89.98874 87.22809 92.85507
## 350 days post coma 90.23479 87.48379 93.09069
## 500 days post coma 90.48151 87.72031 93.34798
## 650 days post coma 90.72890 87.93753 93.62707
## 800 days post coma 90.97697 88.13596 93.92746
```

Part C

i. Explain why having multiple measurements for individuals is useful.

In this case, having multiple measurements is useful because it can benefit the time aspect of the problem. If we examine the IQ of a subject 50 days post coma and 500 days post coma, we could increase our chances of finding correlation between IQ score and time since coma. By doing this, you'd reduce the randomness of changes in IQ scores between time frames. For example, if subject A is tested 50 days post coma and 500 days post coma, we narrow the expected IQ score range since the subject had been previously tested and a already logged a score (score was 88 50 days post coma, we estimate it to be 90 500 days post coma). In contrast, if was test subject A 50 days post coma and subject B 500 days post coma and find a large difference in IQ score, that could be due to the fact that subject B is just more intelligent than subject A. This could lead to inaccurate conclusions that IQ recovers over time because it leaves too much of the test up to the individual rather than the days post coma.

ii. Explain why ignoring this structure could cause problems in the analysis you did.

On the flip side, examining individuals multiple times can also cause problems by creating a generalization based on some of the individuals in the test. An example for this would be if we measured 40 individuals twice out of 100. After our analysis, we find that IQ recovers over time. However, out of the 40 individuals that were measured twice, none of them were in a coma for longer than 30 days. My point here is that when observing multiple individuals twice, you could be reducing the effect that the other variables in the study have on the output (i,e duration of coma and gender) that change when observing different individuals.

Problem 2

Part A

Step 1 - Variable Selection To start, I tried running an automated model selection on the entire Ames housing data set. I ditched this approach quickly because it made my computer crash and also thought that an intuitive approach would be better. When selecting the variables to consider from the start, I tried to put myself in the shoes of someone looking to buy a home. What would I want to know about the house? Total number of rooms? Absolutely. The slope of the property? Probably not.

Below are the starting variables I chose to consider in my model

Step 2 - Take the log of sale price and fit the full model

First, I chose to take a log transformation of the sale price to normalize the distribution of the data as best as possible. Secondly, I fit a full model that considered all variables listed above to start narrowing down what I wanted to keep/ditch/join etc…

```
amesData <- make_ames()
smallames <- select(amesData
,Sale_Price
,MS_Zoning
,Lot_Area
,Street
,Neighborhood
,House_Style
,Overall_Qual
,Exter_Qual
,Year_Built
,Total_Bsmt_SF
,First_Flr_SF
,Second_Flr_SF
,Bsmt_Full_Bath
,Bsmt_Half_Bath
,Full_Bath
,Half_Bath
,TotRms_AbvGrd
,Wood_Deck_SF
,Garage_Area)

smallames$log.Sale_Price <- log(smallames$Sale_Price, 10)
m1 <- lm(log.Sale_Price ~ .-Sale_Price, smallames)
summary(m1)

## 
## Call:
## lm(formula = log.Sale_Price ~ . - Sale_Price, data = smallames)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.85312 -0.02742  0.00255  0.03278  0.32998 
## 
## Coefficients:
##                                                       Estimate Std. Error
## (Intercept)                                          3.066e+00  1.989e-01
## MS_ZoningResidential_High_Density                   -6.742e-03  1.758e-02
## MS_ZoningResidential_Low_Density                     2.083e-02  1.157e-02
## MS_ZoningResidential_Medium_Density                 -8.513e-03  1.318e-02
## MS_ZoningA_agr                                      -2.719e-01  6.610e-02
## MS_ZoningC_all                                      -7.641e-02  1.985e-02
## MS_ZoningI_all                                      -4.113e-03  5.029e-02
## Lot_Area                                             1.018e-06  1.797e-07
## StreetPave                                           3.462e-02  2.064e-02
## NeighborhoodCollege_Creek                            1.072e-02  6.393e-03
## NeighborhoodOld_Town                                -2.781e-03  8.686e-03
## NeighborhoodEdwards                                 -3.464e-02  5.788e-03
## NeighborhoodSomerset                                 3.852e-02  1.174e-02
## NeighborhoodNorthridge_Heights                       4.939e-02  8.321e-03
## NeighborhoodGilbert                                  9.209e-03  7.426e-03
## NeighborhoodSawyer                                  -4.438e-03  6.143e-03
## NeighborhoodNorthwest_Ames                           5.848e-03  6.891e-03
## NeighborhoodSawyer_West                             -1.365e-02  7.378e-03
## NeighborhoodMitchell                                 4.889e-04  7.161e-03
## NeighborhoodBrookside                                1.510e-02  8.749e-03
## NeighborhoodCrawford                                 7.754e-02  7.456e-03
## NeighborhoodIowa_DOT_and_Rail_Road                  -5.647e-03  1.114e-02
## NeighborhoodTimberland                               3.032e-02  9.246e-03
## NeighborhoodNorthridge                               4.287e-02  1.001e-02
## NeighborhoodStone_Brook                              5.541e-02  1.108e-02
## NeighborhoodSouth_and_West_of_Iowa_State_University -1.220e-02  1.079e-02
## NeighborhoodClear_Creek                              4.062e-02  1.062e-02
## NeighborhoodMeadow_Village                          -6.211e-02  1.352e-02
## NeighborhoodBriardale                               -6.782e-02  1.427e-02
## NeighborhoodBloomington_Heights                      9.562e-03  1.360e-02
## NeighborhoodVeenker                                  2.146e-02  1.408e-02
## NeighborhoodNorthpark_Villa                         -2.999e-02  1.430e-02
## NeighborhoodBlueste                                 -1.445e-02  2.180e-02
## NeighborhoodGreens                                   2.447e-03  2.377e-02
## NeighborhoodGreen_Hills                              2.274e-01  4.620e-02
## NeighborhoodLandmark                                -1.818e-02  6.580e-02
## House_StyleOne_and_Half_Unf                         -1.005e-02  1.580e-02
## House_StyleOne_Story                                 3.240e-03  6.301e-03
## House_StyleSFoyer                                    1.361e-02  9.534e-03
## House_StyleSLvl                                      4.570e-03  7.833e-03
## House_StyleTwo_and_Half_Fin                          4.472e-02  2.416e-02
## House_StyleTwo_and_Half_Unf                          3.619e-03  1.447e-02
## House_StyleTwo_Story                                -9.776e-03  5.596e-03
## Overall_QualPoor                                    -1.444e-02  5.004e-02
## Overall_QualFair                                     1.448e-01  4.816e-02
## Overall_QualBelow_Average                            2.168e-01  4.745e-02
## Overall_QualAverage                                  2.659e-01  4.747e-02
## Overall_QualAbove_Average                            3.004e-01  4.759e-02
## Overall_QualGood                                     3.247e-01  4.772e-02
## Overall_QualVery_Good                                3.664e-01  4.799e-02
## Overall_QualExcellent                                4.207e-01  4.871e-02
## Overall_QualVery_Excellent                           3.846e-01  5.039e-02
## Exter_QualFair                                      -8.207e-02  1.593e-02
## Exter_QualGood                                      -3.058e-02  9.374e-03
## Exter_QualTypical                                   -4.973e-02  1.016e-02
## Year_Built                                           8.053e-04  9.869e-05
## Total_Bsmt_SF                                        2.412e-05  5.174e-06
## First_Flr_SF                                         1.003e-04  7.364e-06
## Second_Flr_SF                                        1.047e-04  8.751e-06
## Bsmt_Full_Bath                                       2.884e-02  2.636e-03
## Bsmt_Half_Bath                                       1.601e-02  5.136e-03
## Full_Bath                                            8.643e-03  3.604e-03
## Half_Bath                                            1.060e-02  3.631e-03
## TotRms_AbvGrd                                       -8.238e-04  1.388e-03
## Wood_Deck_SF                                         5.466e-05  1.030e-05
## Garage_Area                                          7.664e-05  7.742e-06
##                                                     t value Pr(>|t|)    
## (Intercept)                                          15.417  < 2e-16 ***
## MS_ZoningResidential_High_Density                    -0.383 0.701428    
## MS_ZoningResidential_Low_Density                      1.800 0.071984 .  
## MS_ZoningResidential_Medium_Density                  -0.646 0.518539    
## MS_ZoningA_agr                                       -4.113 4.01e-05 ***
## MS_ZoningC_all                                       -3.849 0.000121 ***
## MS_ZoningI_all                                       -0.082 0.934826    
## Lot_Area                                              5.666 1.61e-08 ***
## StreetPave                                            1.677 0.093558 .  
## NeighborhoodCollege_Creek                             1.677 0.093599 .  
## NeighborhoodOld_Town                                 -0.320 0.748855    
## NeighborhoodEdwards                                  -5.985 2.44e-09 ***
## NeighborhoodSomerset                                  3.280 0.001052 ** 
## NeighborhoodNorthridge_Heights                        5.936 3.27e-09 ***
## NeighborhoodGilbert                                   1.240 0.215026    
## NeighborhoodSawyer                                   -0.723 0.470016    
## NeighborhoodNorthwest_Ames                            0.849 0.396201    
## NeighborhoodSawyer_West                              -1.851 0.064324 .  
## NeighborhoodMitchell                                  0.068 0.945576    
## NeighborhoodBrookside                                 1.726 0.084544 .  
## NeighborhoodCrawford                                 10.400  < 2e-16 ***
## NeighborhoodIowa_DOT_and_Rail_Road                   -0.507 0.612301    
## NeighborhoodTimberland                                3.279 0.001054 ** 
## NeighborhoodNorthridge                                4.281 1.92e-05 ***
## NeighborhoodStone_Brook                               5.000 6.09e-07 ***
## NeighborhoodSouth_and_West_of_Iowa_State_University  -1.131 0.258170    
## NeighborhoodClear_Creek                               3.826 0.000133 ***
## NeighborhoodMeadow_Village                           -4.593 4.56e-06 ***
## NeighborhoodBriardale                                -4.754 2.10e-06 ***
## NeighborhoodBloomington_Heights                       0.703 0.482074    
## NeighborhoodVeenker                                   1.525 0.127430    
## NeighborhoodNorthpark_Villa                          -2.097 0.036086 *  
## NeighborhoodBlueste                                  -0.663 0.507501    
## NeighborhoodGreens                                    0.103 0.918014    
## NeighborhoodGreen_Hills                               4.922 9.04e-07 ***
## NeighborhoodLandmark                                 -0.276 0.782353    
## House_StyleOne_and_Half_Unf                          -0.636 0.524918    
## House_StyleOne_Story                                  0.514 0.607128    
## House_StyleSFoyer                                     1.427 0.153673    
## House_StyleSLvl                                       0.583 0.559691    
## House_StyleTwo_and_Half_Fin                           1.851 0.064328 .  
## House_StyleTwo_and_Half_Unf                           0.250 0.802579    
## House_StyleTwo_Story                                 -1.747 0.080764 .  
## Overall_QualPoor                                     -0.289 0.772919    
## Overall_QualFair                                      3.006 0.002668 ** 
## Overall_QualBelow_Average                             4.568 5.12e-06 ***
## Overall_QualAverage                                   5.601 2.34e-08 ***
## Overall_QualAbove_Average                             6.313 3.16e-10 ***
## Overall_QualGood                                      6.804 1.24e-11 ***
## Overall_QualVery_Good                                 7.635 3.06e-14 ***
## Overall_QualExcellent                                 8.636  < 2e-16 ***
## Overall_QualVery_Excellent                            7.632 3.12e-14 ***
## Exter_QualFair                                       -5.152 2.75e-07 ***
## Exter_QualGood                                       -3.262 0.001120 ** 
## Exter_QualTypical                                    -4.896 1.03e-06 ***
## Year_Built                                            8.160 4.97e-16 ***
## Total_Bsmt_SF                                         4.662 3.27e-06 ***
## First_Flr_SF                                         13.622  < 2e-16 ***
## Second_Flr_SF                                        11.958  < 2e-16 ***
## Bsmt_Full_Bath                                       10.941  < 2e-16 ***
## Bsmt_Half_Bath                                        3.117 0.001844 **
## Full_Bath                                             2.398 0.016549 *
## Half_Bath                                             2.920 0.003533 **
## TotRms_AbvGrd                                        -0.593 0.552995
## Wood_Deck_SF                                          5.305 1.22e-07 ***
## Garage_Area                                           9.900  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06413 on 2864 degrees of freedom
## Multiple R-squared:  0.8716, Adjusted R-squared:  0.8687
## F-statistic: 299.2 on 65 and 2864 DF,  p-value: < 2.2e-16
```

Step 3 - Variable Alteration

i - Square Footage Variable

I chose to combine all the square footage variables to create a total square footage variable. By doing this, we can gain a better insight of the overall effect that an increase in square footage has on selling price.

ii - Deck Variable

When estimating the price of a house, I find it hard to imagine that examining the deck from a square footage perspective would be more predictive than examining on the basis of having a deck or not. In other words, I think we'd be able to estimate the log selling price better based on if the house has a deck or not, not whether the deck is 200 or 250 square feet

iii - Garage Variable

Same thought process for this variable as well. I think we'd be able to estimate the log selling price better based on if the house has a garage or not, not whether the garage is 300 or 350 square feet. I understand that it holds more importance than an increase in square footage for the deck (someone with a big truck may need 100 more square feet of space), but the point here is to make a generalization.

iv - Bathroom Variable

Here I chose to ignore the distinction between half-bathrooms and whether the bathroom is in the basement or not. For my model, I want to see the effect that adding a bathroom, regardless of type or location, has on the log selling price.

```r
# Combine square footage variables to create a Total square footage predictor
smallames$Total_SF <- (smallames$Total_Bsmt_SF + smallames$First_Flr_SF +
                         smallames$Second_Flr_SF)
```

```
# Create factor for whether the house has a deck
smallames$Wood_Deck <- as.factor(ifelse(smallames$Wood_Deck_SF==0, "No", 
"Yes"))

# Create factor for whether the house has a garage
smallames$Garage <- as.factor(ifelse(smallames$Garage_Area==0, "No", "Yes"))

# Create total bathroom variable
smallames$Total_Bathroom <- (smallames$Bsmt_Full_Bath + 
smallames$Bsmt_Half_Bath
                             + smallames$Full_Bath + smallames$Half_Bath)
```

Step 4 - Fit new model based on new variables, eliminate variables with collinearity/ambiguity

First, I created a new data set that excluded the variables that I had joined and then fit a model containing all variables. After doing this, I chose to remove exterior quality and house style from the model. My reason was due to inaccurate coefficients and possible collinearity between the two variables and other variables. For instance, if the exterior quality of the house is good, then the overall quality is most likely going to be good as well. For house style, if we have a two story house, odds are pretty high that it will have a greater total square footage than a house with one level.

The other two variables that I excluded (MS_Zoning and Street) seemed too ambiguous to include within the model. The variables did have significant predictive power on estimating log selling price, but for the sake of keeping the model reasonable and simple, I chose to exclude them.

```
# Create new data set excluding the variables that were combined
smallames2 <- select(smallames, -Total_Bsmt_SF, -First_Flr_SF, -
Second_Flr_SF,
                     -Bsmt_Full_Bath, -Bsmt_Half_Bath, -Full_Bath, -
Half_Bath,
                     -Wood_Deck_SF, -Garage_Area, -Sale_Price)
# Fit new model(s)
m2 <- lm(log.Sale_Price ~ ., smallames2)

m3 <- lm(log.Sale_Price ~ . -Exter_Qual -House_Style, smallames2)

# Fit final model
m4 <- lm(log.Sale_Price ~ 
Lot_Area+Overall_Qual+Year_Built+TotRms_AbvGrd+Total_SF
                          +Wood_Deck+Garage+Total_Bathroom, smallames2)
summary(m4)

## 
## Call:
## lm(formula = log.Sale_Price ~ Lot_Area + Overall_Qual + Year_Built + 
##     TotRms_AbvGrd + Total_SF + Wood_Deck + Garage + Total_Bathroom, 
##     data = smallames2)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00802 -0.03185  0.00341  0.03683  0.29629
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.672e+00  1.247e-01  21.429  < 2e-16 ***
## Lot_Area                     1.765e-06  1.803e-07   9.790  < 2e-16 ***
## Overall_QualPoor             9.779e-02  4.098e-02   2.386  0.01708 *
## Overall_QualFair             2.688e-01  3.761e-02   7.149 1.10e-12 ***
## Overall_QualBelow_Average    3.309e-01  3.625e-02   9.127  < 2e-16 ***
## Overall_QualAverage          3.903e-01  3.612e-02  10.807  < 2e-16 ***
## Overall_QualAbove_Average    4.248e-01  3.621e-02  11.730  < 2e-16 ***
## Overall_QualGood             4.695e-01  3.639e-02  12.903  < 2e-16 ***
## Overall_QualVery_Good        5.391e-01  3.667e-02  14.701  < 2e-16 ***
## Overall_QualExcellent        6.162e-01  3.735e-02  16.499  < 2e-16 ***
## Overall_QualVery_Excellent 5.818e-01  3.938e-02  14.773  < 2e-16 ***
## Year_Built                   8.998e-04  6.147e-05  14.637  < 2e-16 ***
## TotRms_AbvGrd                3.178e-03  1.176e-03   2.703  0.00692 **
## Total_SF                     6.808e-05  2.982e-06  22.828  < 2e-16 ***
## Wood_DeckYes                 1.147e-02  2.841e-03   4.037 5.55e-05 ***
## GarageYes                    6.569e-02  6.195e-03  10.604  < 2e-16 ***
## Total_Bathroom               2.465e-02  1.989e-03  12.396  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07156 on 2913 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.8366
## F-statistic: 938.2 on 16 and 2913 DF,  p-value: < 2.2e-16
```

Part B

To understand the effect that the variables have on median selling price, we need to convert the variables back to the original scale to estimate that effect. Based on the “GarageYes” output below, we can estimate that the median selling price of a house will be increased by 16% if the house has a garage, holding all other considered variables constant.

Note - Poor Quality Coefficient

It doesn’t make much sense that our model estimates the median selling price of a house to increase by 25% if it’s in poor condition. However, after examining the quality variable further, we see that out of the entire dataset, there are only 13 houses that got listed in poor condition. In other words, most houses aren’t going to get listed in “poor condition” so theoretically that variable would not be used frequently.

```
sum.m4 <- summary(m4)
(10^sum.m4$coefficients[,1])
```

```
##                (Intercept)                   Lot_Area
##                 470.065852                   1.000004
##           Overall_QualPoor           Overall_QualFair
##                   1.252531                   1.857160
##  Overall_QualBelow_Average        Overall_QualAverage
##                   2.142169                   2.456392
##  Overall_QualAbove_Average           Overall_QualGood
##                   2.659456                   2.947705
##      Overall_QualVery_Good      Overall_QualExcellent
##                   3.460079                   4.132794
## Overall_QualVery_Excellent                 Year_Built
##                   3.817482                   1.002074
##              TotRms_AbvGrd                   Total_SF
##                   1.007344                   1.000157
##               Wood_DeckYes                  GarageYes
##                   1.026765                   1.163296
##             Total_Bathroom
##                   1.058411
```

```
table(smallames2$Overall_Qual)
```

```
## 
##      Very_Poor           Poor           Fair  Below_Average        Average 
##              4             13             40            226            825 
##  Above_Average           Good      Very_Good      Excellent Very_Excellent 
##            732            602            350            107             31
```

Problem 3

Part A

Step 1 - Build off of the simple model

To build my complex model, I chose to just dig deeper into the variables that I considered for my simple model. Below is my analysis and thought process on the interactive and polynomial terms that I implemented into the model.

Step 2 - Add interaction terms

i - interaction between year house was built and overall quality

I figured that this would be a good interaction to throw into the model because intuitively, you would think that the effect that the overall quality has on the log sale price would be significantly dependent on the year the house was built. After implementing this interaction into the model, we see that the interaction effect on log sales price is significant

ii – interaction between year house was built and total square footage

For this interaction, I assumed that the the effect that the year the house was built had on the log sale price was dependent on the total square footage of the house. We can see how this makes sense because a house built in 2005 will most likely have more square footage than a house built in 1985.

Step 3 - Add polynomial terms

i - Total Rooms above ground

My thought process for this variable was that there is a benefit to adding a room to a house, but to what extent? A given house only has so much square feet. Therefore, if we have house with 30 rooms, but no living room, kitchen, and one bathroom, that would hurt the log sales price of the house. Below is a graph that outlines the polynomial effect of adding a bedroom to any given house.

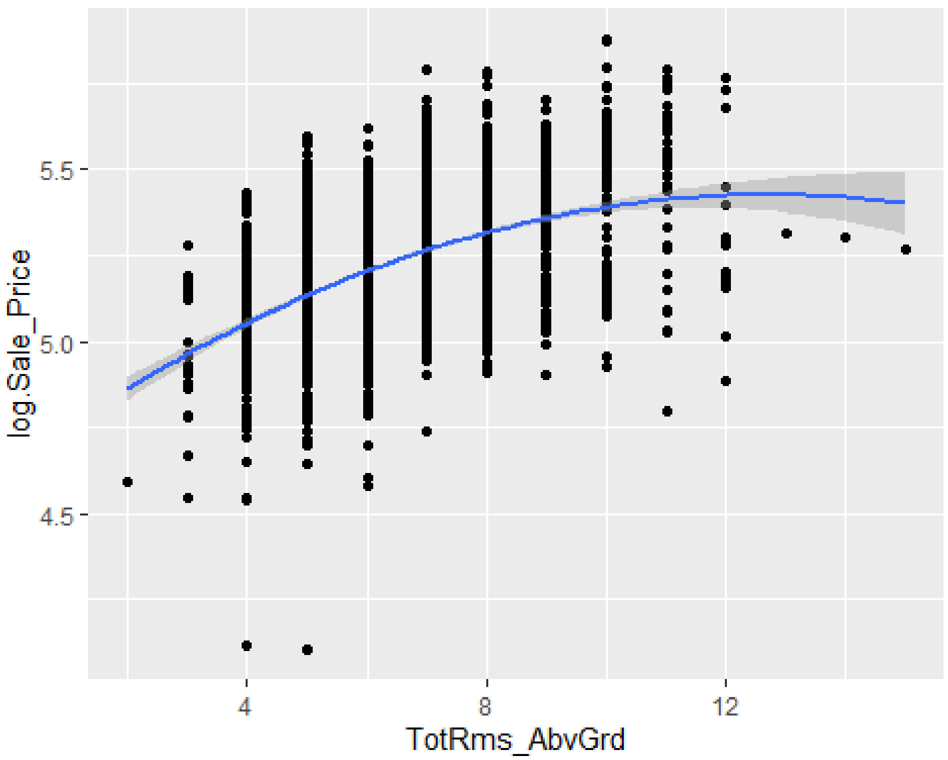
ii - Total Bathrooms

Similar thought process here for the bathrooms in a house. If a house has 15 bathrooms with no bedrooms, kitchen, or living room, that would hurt the log sales price of the house. Below is a graph that also outlines the polynomial effect of adding a bathroom to any given house

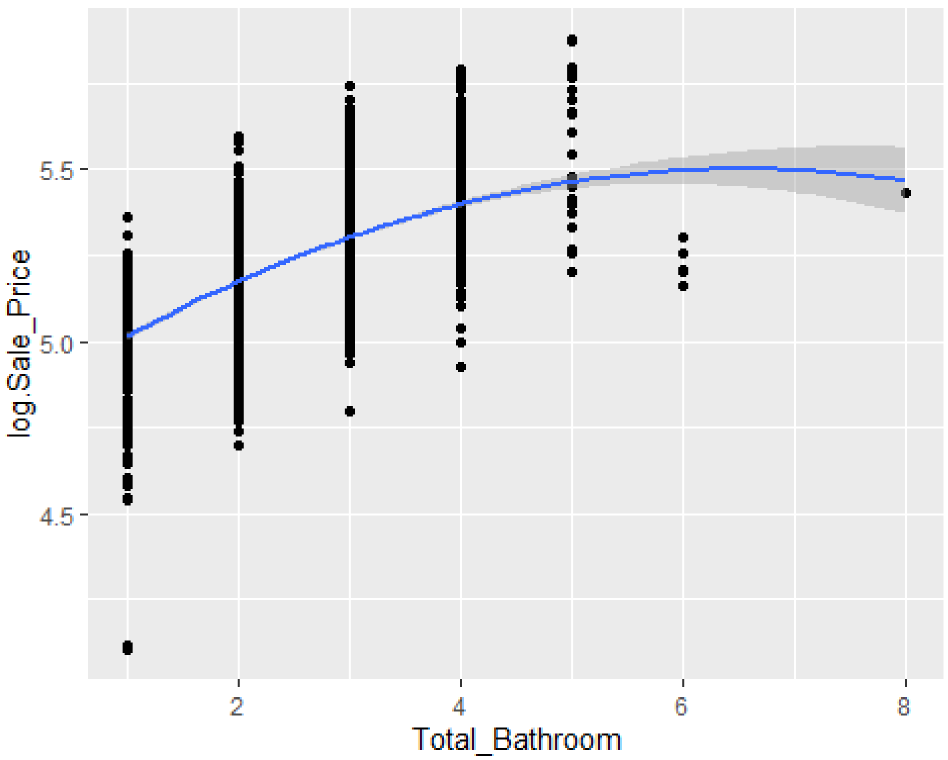
```
p1 <- lm(log.Sale_Price ~ 
Lot_Area+Overall_Qual+Year_Built+TotRms_AbvGrd+I(TotRms_AbvGrd^2)+Total_SF

+Wood_Deck+Garage+Total_Bathroom+I(Total_Bathroom^2)+Year_Built*Overall_Qual
                          +Year_Built*Total_SF, smallames2)
summary(p1)

## 
## Call:
## lm(formula = log.Sale_Price ~ Lot_Area + Overall_Qual + Year_Built + 
##     TotRms_AbvGrd + I(TotRms_AbvGrd^2) + Total_SF + Wood_Deck + 
##     Garage + Total_Bathroom + I(Total_Bathroom^2) + Year_Built * 
##     Overall_Qual + Year_Built * Total_SF, data = smallames2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.87916 -0.03207  0.00311  0.03628  0.28187 
## 
## Coefficients:
##                                         Estimate Std. Error t value
Pr(>|t|)    
## (Intercept)                            2.272e+01  5.580e+00   4.071 4.80e-
05
## Lot_Area                               1.755e-06  1.766e-07   9.936  < 2e-
16
## Overall_QualPoor                      -2.886e+01  6.380e+00  -4.524 6.30e-
06
```

```
## Overall_QualFair                      -2.423e+01  5.692e+00  -4.257 2.13e-
05
## Overall_QualBelow_Average             -2.193e+01  5.588e+00  -3.925 8.88e-
05
## Overall_QualAverage                   -2.131e+01  5.578e+00  -3.821 
0.000136
## Overall_QualAbove_Average             -2.116e+01  5.579e+00  -3.792 
0.000152
## Overall_QualGood                      -2.099e+01  5.580e+00  -3.762 
0.000172
## Overall_QualVery_Good                 -2.138e+01  5.591e+00  -3.823 
0.000134
## Overall_QualExcellent                 -3.384e+01  6.286e+00  -5.383 7.90e-
08
## Overall_QualVery_Excellent            -1.884e+01  5.659e+00  -3.329 
0.000884
## Year_Built                            -9.481e-03  2.872e-03  -3.301 
0.000977
## TotRms_AbvGrd                          3.794e-02  4.906e-03   7.733 1.43e-
14
## I(TotRms_AbvGrd^2)                    -2.544e-03  3.409e-04  -7.464 1.11e-
13
## Total_SF                               7.121e-04  1.478e-04   4.817 1.53e-
06
## Wood_DeckYes                           1.239e-02  2.774e-03   4.468 8.18e-
06
## GarageYes                              6.181e-02  6.108e-03  10.120  < 2e-
16
## Total_Bathroom                         3.719e-02  5.865e-03   6.340 2.66e-
10
## I(Total_Bathroom^2)                   -2.280e-03  9.750e-04  -2.338 
0.019449
## Overall_QualPoor:Year_Built            1.493e-02  3.288e-03   4.541 5.83e-
06
## Overall_QualFair:Year_Built            1.261e-02  2.930e-03   4.303 1.74e-
05
## Overall_QualBelow_Average:Year_Built   1.145e-02  2.876e-03   3.979 7.08e-
05
## Overall_QualAverage:Year_Built         1.115e-02  2.871e-03   3.885 
0.000105
## Overall_QualAbove_Average:Year_Built   1.109e-02  2.872e-03   3.862 
0.000115
## Overall_QualGood:Year_Built            1.103e-02  2.872e-03   3.841 
0.000125
## Overall_QualVery_Good:Year_Built       1.126e-02  2.877e-03   3.914 9.29e-
05
## Overall_QualExcellent:Year_Built       1.752e-02  3.215e-03   5.451 5.44e-
08
## Overall_QualVery_Excellent:Year_Built  1.003e-02  2.911e-03   3.445 
0.000580
```

```
## Year_Built:Total_SF                   -3.250e-07  7.482e-08  -4.343 1.45e-
05
##
## (Intercept)                           ***
## Lot_Area                              ***
## Overall_QualPoor                      ***
## Overall_QualFair                      ***
## Overall_QualBelow_Average             ***
## Overall_QualAverage                   ***
## Overall_QualAbove_Average             ***
## Overall_QualGood                      ***
## Overall_QualVery_Good                 ***
## Overall_QualExcellent                 ***
## Overall_QualVery_Excellent            ***
## Year_Built                            ***
## TotRms_AbvGrd                         ***
## I(TotRms_AbvGrd^2)                    ***
## Total_SF                              ***
## Wood_DeckYes                          ***
## GarageYes                             ***
## Total_Bathroom                        ***
## I(Total_Bathroom^2)                   *
## Overall_QualPoor:Year_Built           ***
## Overall_QualFair:Year_Built           ***
## Overall_QualBelow_Average:Year_Built  ***
## Overall_QualAverage:Year_Built        ***
## Overall_QualAbove_Average:Year_Built  ***
## Overall_QualGood:Year_Built           ***
## Overall_QualVery_Good:Year_Built      ***
## Overall_QualExcellent:Year_Built      ***
## Overall_QualVery_Excellent:Year_Built ***
## Year_Built:Total_SF                   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06977 on 2901 degrees of freedom
## Multiple R-squared:  0.8462, Adjusted R-squared:  0.8447
## F-statistic: 569.8 on 28 and 2901 DF,  p-value: < 2.2e-16

# Visualization of the polynomial terms
p <- ggplot(smallames2, aes(x = TotRms_AbvGrd, y =
log.Sale_Price))+geom_point()
q1 <- lm(log.Sale_Price ~ poly(TotRms_AbvGrd,2), smallames2)
p+geom_smooth(method="lm", formula=y~poly(x,2))
```



```
p_2 <- ggplot(smallames2, aes(x = Total_Bathroom, y =
log.Sale_Price))+geom_point()
q2 <- lm(log.Sale_Price ~ poly(Total_Bathroom,2), smallames2)
p_2+geom_smooth(method="lm", formula=y~poly(x,2))
```



Part B

Coefficient Confusion

In assignment 9 when I was working through this problem with a simpler model and I added an interaction term, the standalone coefficients made more sense, but the interaction coefficient was slightly puzzling. Here we have the opposite case - the standalone coefficients that were put into an interaction term (Year_Built, Overall_Qual, and Total_SF) are puzzling, but the interaction coefficients make more sense.

```
sum.p1 <- summary(p1)
data.frame(sum.p1$coefficients[,1])

##                                       sum.p1.coefficients...1.
## (Intercept)                                       2.271985e+01
## Lot_Area                                          1.754683e-06
## Overall_QualPoor                                 -2.886419e+01
## Overall_QualFair                                 -2.423244e+01
## Overall_QualBelow_Average                        -2.193145e+01
## Overall_QualAverage                              -2.131063e+01
## Overall_QualAbove_Average                        -2.115785e+01
## Overall_QualGood                                 -2.099362e+01
## Overall_QualVery_Good                            -2.137603e+01
## Overall_QualExcellent                            -3.384103e+01
## Overall_QualVery_Excellent                       -1.883729e+01
## Year_Built                                       -9.480547e-03
## TotRms_AbvGrd                                     3.793936e-02
## I(TotRms_AbvGrd^2)                               -2.544248e-03
## Total_SF                                          7.120839e-04
## Wood_DeckYes                                      1.239434e-02
## GarageYes                                         6.181356e-02
## Total_Bathroom                                    3.718562e-02
## I(Total_Bathroom^2)                              -2.279637e-03
## Overall_QualPoor:Year_Built                       1.492875e-02
## Overall_QualFair:Year_Built                       1.260951e-02
## Overall_QualBelow_Average:Year_Built              1.144525e-02
## Overall_QualAverage:Year_Built                    1.115333e-02
## Overall_QualAbove_Average:Year_Built              1.109124e-02
## Overall_QualGood:Year_Built                       1.103165e-02
## Overall_QualVery_Good:Year_Built                  1.126141e-02
## Overall_QualExcellent:Year_Built                  1.752225e-02
## Overall_QualVery_Excellent:Year_Built             1.002654e-02
## Year_Built:Total_SF                              -3.249500e-07
```

Problem 3

Simple or Complex Model

If I had to choose between the two models to present in front of group of people, I would choose the simpler model every time. The simpler model accounts for enough of the randomness (Adj. R-squared of 84%) in estimating log selling price and it is much easier to interpret. The complex model is much more specific and possibly more precise regarding each variable in the model (i.e polynomial and interactive terms), but I can't fully grasp why the output is the way it is. Along with this, the complex model doesn't explain any more of the randomness in estimating the log selling price of a house (Adj. R-squared of 84%). I can't really see my preference leaning towards the complex model in any scenario because it's so hard to interpret. I don't have much use for a model where I can't explain the estimated output and how it came to be.