Pooja Deshpande

Computational Assignment 2
MSDS – 410 Data Modeling for Supervised Learning, Summer 2020
Northwestern University

Contents:

PART. I

Anova table:

ANOVA:					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1974.53	1974.53	209.8340	< 0.0001
X2	1	118.8642568	118.8642568	12.6339	0.0007
X3	1	32.47012585	32.47012585	3.4512	0.0676
X4	1	0.435606985	0.435606985	0.0463	0.8303
Residuals	67	630.36	9.41		
Note: You can make the following calculations from the ANOVA table above to get Overall F statistic					
Model (adding 4 rows)	4	2126	531.50		<0.0001
Total (adding all rows)	71	2756.37			

Coefficients:				
	Estimate	Std. Error	t value	Pr(>t)
Intercept	11.3303	1.9941	5.68	<.0001
X1	2.186	0.4104		<.0001
X2	8.2743	2.3391	3.54	0.0007
X3	0.49182	0.2647	1.86	0.0676
X4	-0.49356	2.2943	-0.22	0.8303

Residual standard error: 3.06730 on 67 degrees of freedom	
Multiple R-sqaured: 0.7713, Adjusted R-squared: 0.7577	
F-statistic: on 4 and 67 DF, p-value < 0.0001	

Number of predictors	C(p)	R-square	AIC	BIC	Variables in the model
4	5	0.7713	166.2129	168.9481	X1 X2 X3 X4

1. How many observations are in the sample data?

The formula for degrees of freedom of residuals is n-p-1, where n is the sample size, p is the number of estimated parameters.

df = n -p -1

67 = n – 4 -1

n = 67 + 5

n = 72

2. Write out the null and alternate hypotheses for the t-test for Beta1.

Null Hypothesis, $H_{0:}$ $\beta 1 = 0$

Alternate Hypothesis, $H_{a:}$ $\beta 1 \neq 0$

3. Compute the t- statistic for Beta1. Conduct the hypothesis test and interpret the result.

t = $\beta 1\_hat/s.e\ (\beta 1\_hat)$

t = 2.186/0.4104

t = 5.3265

```
> p_value <- pt(5.3265,70,lower.tail = FALSE)*2
> p_value
[1] 0.000001153488
> critical_t_value <- qt(0.025,70,lower.tail = FALSE)
> critical_t_value
[1] 1.994437
```

p-value is lower than the significance level, alpha = 0.05

Critical t – value in this case is 1.9944 which is lesser than the result from the t- test. Hence, we can reject the null hypothesis that there is no significant relationship between X1 and Y.

4. Compute the R-Squared value for Model 1, using information from the ANOVA table. Interpret this statistic.

R-Square = 1 – Sum of squares of residuals/Total sum of squares

= 1 – (630.36)/ (1974.53+118.8642 + 32.4701 + 0.4356 + 630.36)

= 1 – (630.36)/ (2756.6599)

= 1 – 0.2287 = 0.7713

77% of the variance in the response variable is explained by X1, X2, X3, and X4.

5. Compute the Adjusted R-Squared value for Model 1. Discuss why Adjusted R-squared and the R-squared values are different.

Adjusted R square = 1 – ((1 – R2) * (n – 1)/ (n - p – 1))

Where R2 is R square, n is sample size, p is number of estimated parameters.

Adjusted R square = 1 – ((1 – 0.7713) * (72 – 1)/ (72 - 4 – 1))

= 1 – 0.2423

= 0.7576

R square value always increases as the more predictors are added to the model. Adjusted R square penalizes the model for every new predictor that gets added. Adjusted R square increases only when the new term contributes the model more than expected by chance.

6. Write out the null and alternate hypotheses for the Overall F-test.

The null and alternate hypotheses for the overall F- test can be written in two ways.

Reduced model: $H_{0:}$ $Y = \beta 0$

Full model: $H_{a:}$ $Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \varepsilon$

(or)

$H_{0:}$ $\beta 1 = \beta 2 = \beta 3 = \beta 4 = 0$

$H_{a:}$ At least one of the parameters is not equal to zero.

7. Compute the F-statistic for the Overall F-test. Conduct the hypothesis test and interpret the result.

F = (SSR/p)/ (SSE/(n – p – 1))

F = MSR/MSE

F = (2126.2999/4)/(630.36/67)

F = 531.5749/9.4083

F = 56.5

```
> #Q7
> critical_f_value <- qf(0.05,4,67,lower.tail = FALSE)
> critical_f_value
[1] 2.508695
> p_value <- pf(56.5,4,67,lower.tail = FALSE)
> p_value
[1] 0.00000000000000000000009166773
> round(p_value ,4)
[1] 0
```

Critical F value is 2.5087. F-test statistic of 56.5 is greater than the critical value.

P-value of the test is 0. It is lower than the significance level of alpha = 0.05.

Hence, we can reject the null hypothesis that all the parameters are equal to zero. There is at least one parameter which has a significant relationship with the response variable.

ANOVA:					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X1	1	1928.27000	1928.27000	218.8890	<.0001
X2	1	136.92075	136.92075	15.5426	0.0002
X3	1	40.75872	40.75872	4.6267	0.0352
X4	1	0.16736	0.16736	0.0190	0.8908
X5	1	54.77667	54.77667	6.2180	0.0152
X6	1	22.86647	22.86647	2.5957	0.112
Residuals	65	572.60910	8.80937		
Note: You can make the following calculations from the ANOVA table above to get Overall F statistic					
Model (adding 6 rows)	6	2183.75946	363.96	41.3200	<0.0001
Total (adding all rows)	71	2756.37			

Coefficients:				
	Estimate	Std. Error	t value	Pr(>t)
Intercept	14.3902	2.89157	4.98	<.0001
X1	1.97132	0.43653	4.52	<.0001
X2	9.13895	2.30071	3.97	0.0002
X3	0.56485	0.26266	2.15	0.0352
X4	0.33371	2.42131	0.14	0.8908
X5	1.90698	0.76459	2.49	0.0152
X6	-1.0433	0.64759	-1.61	0.112
Residual standard error: 2.968 on 65 degrees of freedom				
Multiple R-sqaured: 0.7923, Adjusted R-squared: 0.7731				
F-statistic: 41.32 on 6 and 65 DF, p-value < 0.0001				

Number of predictors	C(p)	R-square	AIC	BIC	Variables in the model
6	7	0.7923	163.2947	166.7792	X1 X2 X3 X4 X5 X6

8. Now let's consider Model 1 and Model 2 as a pair of models. Does Model 1 nest Model 2 or does Model 2 nest Model 1? Explain.

Two models are nested if one consists of all the terms in the other, and at least one additional term. Model 2 consists of all the terms in Model 1 and two additional terms. Hence Model 1 is nested within Model 2. One can say that the Model 2 is the full model and Model 1 is the reduced model.

Model 2 consists of X1, X2, X3, X4, X5 and X6.

Model 1 consists of X1, X2, X3 and X4.

9. Write out the null and alternate hypotheses for a nested F-test using Model 1 and Model 2.

RM: Y = $\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \varepsilon$

FM: $Y = \beta 0 + \beta X11 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \beta 6X6 + \varepsilon$

H_o: Reduced model is adequate

H_a: Full model is adequate

(or)

H_o: $\beta 5 = \beta 6 = 0$

10. Compute the F-statistic for a nested F-test using Model 1 and Model 2. Conduct the hypothesis test and interpret the results.

$F = ([SSE(RM) – SSE(FM)]/ (p + 1 – k))/ (SSE(FM)/ (n – p – 1))$

= ([630.36 – 572.6091]/ (6 + 1 – 4))/ (572.6091/ (72 – 6 – 1)

= (57.7509/ 3)/ (572.6091/ 65)

= 19.2503/ 8.8093

= 2.1852

```
> critical_f_value <- qf(0.05,3,65,lower.tail = FALSE)
> critical_f_value
[1] 2.745915
> p_value <- pf(2.1852,3,65,lower.tail = FALSE)
> round(p_value,4)
[1] 0.0982
```

Critical F value is greater than the observed F statistic and the p-value is greater than the significance level. This means that there isn't enough statistical evidence to reject the null hypothesis. From this, one can conclude that the reduced model i.e. the Model 1 is adequate. It is always better to have a simpler model with fewer variables.

PART. 2

11. After considering all the continuous explanatory variables, the following were chosen:

LotArea, TotalBsmtSF, GrLivArea, PoolArea, FirePlaces, GarageArea, WoodDeckSF, OpenPorchSF, TotalSqft, qual_index

TotalSqft is the sum of First floor square feet and second floor square feet.

qual_index is the product of overall condition and overall quality.

These variables can be grouped based on whether they describe features inside the house, outside the house or the entire home.

Group 1(Outside the house): *PoolArea, WoodDeckSF, OpenPorchSF, GarageArea*

Group 2(Inside the house): *TotalBsmtSF, GrLivArea, TotalSqft, Fireplaces (discrete variable)*

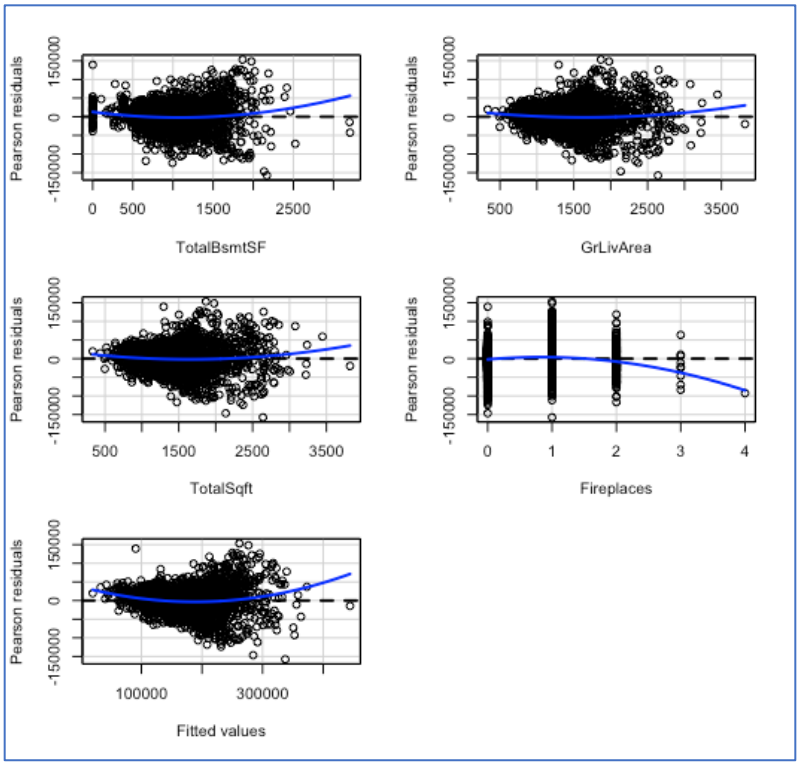
Group 3(Entire home): *LotArea, qual_index (discrete variable)*

12. The first group that will be used to build a model is group 2.

```
> model3 <- lm(SalePrice ~ TotalBsmtSF + GrLivArea + TotalSqft + Fireplaces,data=final_df)
> summary(model3)

Call:
lm(formula = SalePrice ~ TotalBsmtSF + GrLivArea + TotalSqft +
    Fireplaces, data = final_df)

Residuals:
    Min      1Q  Median      3Q     Max
-157469  -19698     882   19515  153234

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) -4983.793   2781.471  -1.792              0.0733 .
TotalBsmtSF    61.649      2.013  30.625 < 0.0000000000000002 ***
GrLivArea     -46.931     18.279  -2.567              0.0103 *
TotalSqft     120.794     18.317   6.595      0.0000000000526 ***
Fireplaces  13611.598   1295.748  10.505 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35450 on 2308 degrees of freedom
Multiple R-squared:  0.6915,    Adjusted R-squared:  0.6909
F-statistic:  1293 on 4 and 2308 DF,  p-value: < 0.00000000000000022

> coef(model3)
(Intercept) TotalBsmtSF   GrLivArea   TotalSqft  Fireplaces
-4983.79279    61.64938   -46.93088   120.79449 13611.59825
```

```
> confint(model3)
                   2.5 %      97.5 %
(Intercept) -10438.23701   470.65143
TotalBsmtSF     57.70188    65.59687
GrLivArea      -82.77665   -11.08510
TotalSqft       84.87530   156.71368
Fireplaces   11070.64588 16152.55061
> anova(model3)
Analysis of Variance Table

Response: SalePrice
             Df        Sum Sq       Mean Sq  F value               Pr(>F)
TotalBsmtSF    1 3540655285243 3540655285243 2817.446 < 0.00000000000000022 ***
GrLivArea      1 2759261758570 2759261758570 2195.659 < 0.00000000000000022 ***
TotalSqft      1   61341423686   61341423686   48.812      0.000000000003672 ***
Fireplaces     1  138677186137  138677186137  110.351 < 0.00000000000000022 ***
Residuals   2308 2900439584072    1256689594
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual plots:



T-test for TotalBsmtSqft:

Ho: $\beta_{TotalBsmtSqft} = 0$

Ha: $\beta_{TotalBsmtSqft} != 0$

p-value < 0.00001, TotalBsmtSqft is a significant predictor. For unit change in TotalBsmtSqft, SalePrice increases by 61.649 when adjusted for all predictors.

T-test for GrLivArea:

Ho: $\beta_{GrLivArea} = 0$

Ha: $\beta_{GrLivArea} != 0$

p-value = 0.0103, GrLivArea is a significant predictor. For unit change in GrLivArea, SalePrice decreases by 46.931 when adjusted for all predictors.

T-test for TotalSqft:

Ho: $\beta_{TotalSqft} = 0$

Ha: $\beta_{TotalSqft} \; != 0$

p-value < 0.00001, TotalSqft is a significant predictor. For unit change in TotalSqft, SalePrice increases by 120.794 when adjusted for all predictors.

T-test for FirePlaces:

Ho: $\beta_{FirePlaces} = 0$

Ha: $\beta_{Fireplaces} \; != 0$

p-value < 0.00001, Fireplaces is a significant predictor. For unit change in Fireplaces, SalePrice increases by 13611.598 when adjusted for all predictors.

Omnibus overall F-test:

Ho: $\beta_{FirePlaces} = \beta_{TotalSqft} = \beta_{TotalBsmtSqft} = \beta_{GrLivArea} = 0$

Ha: At least one of the parameters is not equal to zero.

p-value < 0.00001. At least one of the variables have a significant relationship with the response variable. However, the residual plots indicate non null plots. This means that some of the assumptions may have been violated. One reason could be collinearity.

13. Group 1 variables will be added to model 3. List of variables in Model 4:

TotalBsmtSF (X1), GrLivArea (X2), TotalSqft (X3), Fireplaces (X4), PoolArea (X5), WoodDeckSF (X6), OpenPorchSF (X7), GarageArea (X8).

```
> #Model 4
> model4 <- lm(SalePrice ~ TotalBsmtSF + GrLivArea + TotalSqft + Fireplaces + PoolArea + WoodDeckSF +
+                OpenPorchSF + GarageArea, data = final_df)
> summary(model4)

Call:
lm(formula = SalePrice ~ TotalBsmtSF + GrLivArea + TotalSqft +
    Fireplaces + PoolArea + WoodDeckSF + OpenPorchSF + GarageArea,
    data = final_df)

Residuals:
    Min      1Q  Median      3Q     Max
-189077  -17342     422   17858  147242

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept) -10575.606   2558.830  -4.133 0.000037088228967178 ***
TotalBsmtSF     47.037      1.923  24.465 < 0.0000000000000002 ***
GrLivArea      -32.835     16.496  -1.990              0.04666 *
TotalSqft       91.654     16.572   5.531 0.000000035494849377 ***
Fireplaces   12070.720   1178.564  10.242 < 0.0000000000000002 ***
PoolArea       -56.118     20.994  -2.673              0.00757 **
WoodDeckSF      44.676      5.427   8.232 0.000000000000000305 ***
OpenPorchSF     73.169     11.489   6.369 0.000000000229335098 ***
GarageArea      77.857      4.018  19.378 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31970 on 2304 degrees of freedom
Multiple R-squared:  0.7495,    Adjusted R-squared:  0.7487
F-statistic: 861.9 on 8 and 2304 DF,  p-value: < 0.00000000000000022
```

```
> coef(model4)
 (Intercept)  TotalBsmtSF    GrLivArea    TotalSqft   Fireplaces     PoolArea   WoodDeckSF
-10575.60619     47.03664    -32.83486     91.65406  12070.71968    -56.11783     44.67555
 OpenPorchSF   GarageArea
    73.16897     77.85700
> confint(model4)
                  2.5 %        97.5 %
(Intercept) -15593.45631 -5557.7560626
TotalBsmtSF     43.26635    50.8069304
GrLivArea      -65.18394    -0.4857682
TotalSqft       59.15695   124.1511710
Fireplaces    9759.56306 14381.8762907
PoolArea       -97.28757   -14.9480941
WoodDeckSF      34.03279    55.3183177
OpenPorchSF     50.63957    95.6983722
GarageArea      69.97807    85.7359231
> anova(model4)
Analysis of Variance Table

Response: SalePrice
              Df        Sum Sq       Mean Sq   F value                Pr(>F)
TotalBsmtSF    1 3540655285243 3540655285243 3464.8612 < 0.00000000000000022 ***
GrLivArea      1 2759261758570 2759261758570 2700.1948 < 0.00000000000000022 ***
TotalSqft      1   61341423686   61341423686   60.0283   0.00000000000001391 ***
Fireplaces     1  138677186137  138677186137  135.7086 < 0.00000000000000022 ***
PoolArea       1    3494376003    3494376003    3.4196              0.06456 .
WoodDeckSF     1   98071391498   98071391498   95.9720 < 0.00000000000000022 ***
OpenPorchSF    1   60755915286   60755915286   59.4553   0.00000000000001847 ***
GarageArea     1  383717728292  383717728292  375.5036 < 0.00000000000000022 ***
Residuals   2304 2354400172993    1021875075
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hypothesis tests for individual parameters:

For X1,

H0: $\beta 1 = 0$

Ha: $\beta 1 \neq 0$

p- value < 0.000001 suggests TotalBsmtSF has a statistically significant relationship with SalePrice.

For X2,

H0: $\beta 2 = 0$

Ha: $\beta 2 \neq 0$

p-value = 0.0466 suggests GrLivArea has a statistically significant relationship with SalePrice.

For X3,

H0: $\beta 3 = 0$

Ha: $\beta 3 \neq 0$

p-value < 0.00001 suggests TotalSqft has a statistically significant relationship with SalePrice.

For X4,

H0: $\beta 4 = 0$

Ha: $\beta 4 \neq 0$

p-value < 0.00001 suggests Fireplaces has a statistically significant relationship with SalePrice.

For X5,

H0: $\beta 5 = 0$

Ha: $\beta 5 \neq 0$

p-value = 0.00757 suggests PoolArea has a statistically significant relationship with SalePrice

For X6,

H0: $\beta 6 = 0$

Ha: $\beta 6 \neq 0$

p-value < 0.00001 suggests that WoodDeckSF has a statistically significant relationship with SalePrice.

For X7,

H0: $\beta 7 = 0$

Ha: $\beta 7 \neq 0$

p-value < 0.00001 suggests that OpenPorchSF has a statistically significant relationship with SalePrice.

For X8,

H0: $\beta 8 = 0$

Ha: $\beta 8 \neq 0$

p-value < 0.0001 suggests that GarageArea has a statistically significant relationship with SalePrice.

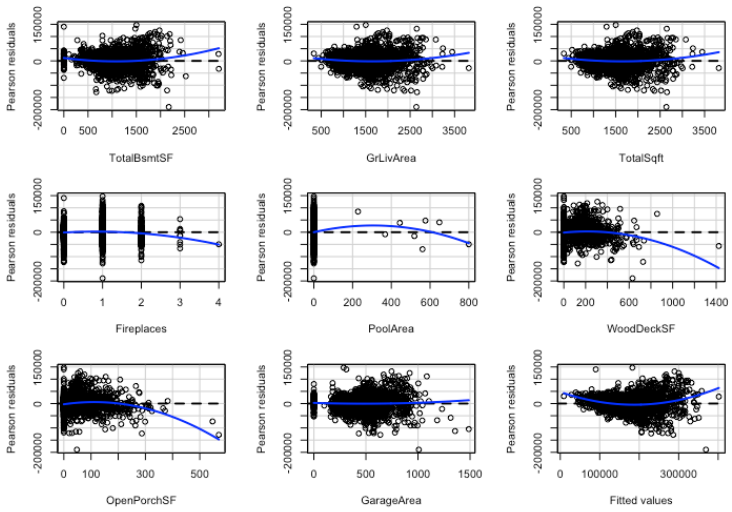
Omnibus overall F- test:

H0: $\beta 1 = \beta 2 = \beta 3 = \beta 4 = \beta 5 = \beta 6 = \beta 7 = \beta 8 = 0$

Ha: At least one of the parameters is not equal to zero.

p-value < 0.00001 suggests that at least one regressor has a significant relationship with SalePrice.

Residual Plots:



14. Nested F-test for Model 3 and Model 4:

RM: Y = $\beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \varepsilon$

FM: Y $= \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5 + \beta 6X6 + \beta 7X7 + \beta 8X8 + \varepsilon$

H_o: Reduced model is adequate

H_a: Full model is adequate

(or)

H_o: $\beta 5 = \beta 6 = \beta 7 = \beta 8 = 0$

$F = ([SSE(RM) – SSE(FM)]/ (p + 1 – k))/ (SSE(FM)/ (n – p – 1))$

$= ([2900439584072 - 2354400172993]/ (8 + 1 – 4))/ (2354400172993/ (2313 – 8 – 1))$

$= (546039411079/5)/ (2354400172993/2304)$

$= (109207882216)/ 1021875075$

= 106.8701

Critical value of F is 2.2179. The observed F-test statistic is way larger than the critical value which suggests that there is enough statistical evidence to reject the null hypothesis.

p-value < 0.000001 suggests that the null hypothesis can be rejected. Full model is adequate. This seems like a strange result to me since model 3 had a significant F-test. This could be a result of assumptions violation.