**Team 19**

**Akshat Sangal**

**Murli Manohar Singh**

**Project 2 Part 2- An Analysis on the Ames Home Sales Data Set**

IDS 462 Statistical Software In Business

**Question 1**:- Write a macro to invoke PROC GLMSELECT five times on the **SalePrice** variable regressing on the interval variables. For each, request STEPWISE selection with the SELECTION= option and include DETAILS=STEPS to obtain step information and the selection summary table. Use 0.05 as the significance level for entry into and staying in the model. Call to macro to run SELECT for the options SL, AIC, BIC, AICC, and SBC and compare the selected models from the output. Does the significance level for entry into and staying in the model have any impact when you use options other than SL? Which variables stay in the model for each 5 options? Which selection methods and criteria would you recommend?

**Answer 1**:-

**Functions Description**:-

* **%LET**

%LET is used to define a macro variable. The syntax for that is:-

**%LET macro-variable =<value>;**

macro-variable is either the name of a macro variable or a text expression that produces a macro variable name. The name can refer to a new or existing macro variable.

**Data Analysis**:-

Q 1: Does the significance level for entry into and staying in the model have any impact when you use options other than SL

A 1: No, other than SL none of the options have any impact of significance level for entry into and staying in the model

Q 2: Which variables stay in the model for each 5 options

A 2: Below are the variables those stay in all the 5 models:

Gr\_Liv\_Area

Basement\_Area

Deck\_Porch\_Area

Age\_Sold

Log\_Price

Q 3: Which selection methods and criteria would you recommend

A 3: All the methods have same r-square value and Root MSE. So, I would suggest to go with any model all will be effective.

* %LET statements are used to name the macro variables and set their values
* “interval” is a macro variable created for all the continuous variables
* “interval1” is a macro variable created for all the continuous variables for this question specifically as we cannot include Sales price in this macro as we are building model on it
* “categorical” is a macro variable created for all the categorical variables
* “interval” macro variable consists of the following continuous variables:-

1. Lot\_Area
2. Gr\_Liv\_Area
3. Garage\_Area
4. SalePrice
5. Basement\_Area
6. Deck\_Porch\_Area
7. Log\_Price
8. Age\_Sold

* “categorical” macro variable consists of the following categorical variables:-

1. House\_Style
2. Overall\_Qual
3. Overall\_Cond
4. Heating\_Qc
5. Central\_Air
6. Bedroom\_AbvGr
7. FirePlaces
8. Mo\_Sold
9. Full\_Bathroom
10. Half\_Bathroom
11. Total\_Bathroom
12. Season\_Sold
13. Garage\_Type\_2
14. Foundation\_2
15. Masonry\_Veneer
16. Lot\_Shape\_2
17. House\_style2
18. Overall\_Qual2
19. Overall\_Cond2
20. Bonus

**Question 2:-** Invoke PROC REG with the plots option using rsquare adjrsq cp to produce a regression of **SalePrice** on all the other interval variables in the data set. Use selection = rsquare adjrsq cp. Which model you would suggest, and why? Hint: compare rsquare adjrsq cp.

**Answer 2**:-

We will use same macro "interval1" in this question too. I would suggest to go with R-Square as it is explaining 95.82% of variation with just 4 variable selection:

Number in Model= 4

R-Square= 0.9585

Adjusted R-Square=0.9582

C(p)=10.6189  
Variables in Model = Gr\_Liv\_Area Basement\_Area Age\_Sold Log\_Price

**Question 3:-** Invoke PROC FREQ and create one‑way frequency tables for the variables **Bonus**, **Fireplaces**, and **Lot\_Shape\_2** and create two‑way frequency tables for the variables **Bonus** by **Fireplaces**, and **Bonus** by **Lot\_Shape\_2**. For the continuous variable, **Basement\_Area**, create histograms for each level of **Bonus**. Use a CLASS statement in PROC UNIVARIATE. Use the FORMAT procedure to format the values of **Bonus**.

Are there any unusual data values that could be due to coding errors for any of the categorical variables?

Examine the distribution of **Bonus** at each value of the predictors. What associations do you see?

**Answer 3**:-

No we have not observed any unusual data values because of coding errors for categorical variables.

However in case of interval variable, there is a bonus associated with greater basement area. So chances of bonus are much more when you have house basement are greater than 1200 square feet.

**Appendix**

This section contains the SAS project code and the screenshots of all the outputs for all the questions in the project. Code description is there in the comments of each code snippet.

**QUESTION 1**

**CODE**

libname proj2 '/folders/myfolders/project2\_1';

%Let categorical=House\_Style Overall\_Qual Overall\_Cond Heating\_QC Central\_Air Bedroom\_AbvGr Fireplaces Mo\_Sold Full\_Bathroom Half\_Bathroom Total\_Bathroom Season\_Sold Garage\_Type\_2 Foundation\_2 Masonry\_Veneer Lot\_Shape\_2 House\_Style2 Overall\_Qual2 Overall\_Cond2 Bonus;

%Let interval=Lot\_Area Gr\_Liv\_Area Garage\_Area SalePrice Basement\_Area Deck\_Porch\_Area Age\_Sold Log\_Price;

/\* Part 2 – Prob 1:

Q 1: Does the significance level for entry into and staying in the model

have any impact when you use options other than SL

A 1: No, other than SL none of the options have any impact of

ignificance level for entry into and staying in the model

Q 2: Which variables stay in the model for each 5 options

A 2: Below are the variables those stay in all the 5 models:

Gr\_Liv\_Area

Basement\_Area

Deck\_Porch\_Area

Age\_Sold

Log\_Price

Q 3: Which selection methods and criteria would you recommend

A 3: All the methods have same r-square value and Root MSE.

So, I would suggest to go with any model all will be effective. \*/

ods graphics on;

%Let interval1=Lot\_Area

Gr\_Liv\_Area

Garage\_Area

Basement\_Area

Deck\_Porch\_Area

Age\_Sold

Log\_Price;

%macro macro\_glm(method);

proc glmselect data=proj2.team19 plots=all;

model SalePrice=&interval1/

SELECTION=STEPWISE DETAILS=steps SELECT=&method SLENTRY=0.05 SLSTAY=0.05;

title "Macro execution with Stepwise Selection and method &method";

run;

quit;

%mend macro\_glm;

%macro\_glm(SL);

%macro\_glm(AIC);

%macro\_glm(BIC);

%macro\_glm(AICC);

%macro\_glm(SBC);

ods graphics off;

**QUESTION 2**

**CODE**

/\* Part 2 – Prob 2: We will use same macro "interval1" in this question too.

I would suggest to go with R-Square as it is expalaining 95.82% of variation

with just 4 variable selection:

Number in R-Square Adjusted R-Square C(p) Variables in Model

Model

4 0.9585 0.9582 10.6189 Gr\_Liv\_Area Basement\_Area Age\_Sold Log\_Price

\*/

proc reg data=proj2.team19;

model SalePrice=&interval1/selection=rsquare adjrsq cp;

run;

**QUESTION 3**

**CODE**

/\* Part 2 – Prob 3:

There is a bonus associsted with greater basment area.

So chances of bonus are much more when you have house basement are greater than 1200 square feet

\*/

ods graphics on;

proc freq data=proj2.team19;

tables Bonus Fireplaces Lot\_Shape\_2 Bonus\*Fireplaces Bonus\*Lot\_Shape\_2;

run;

proc univariate data=proj2.team19;

histogram Basement\_Area;

class Bonus;

var Basement\_Area;

run;

ods graphics off;