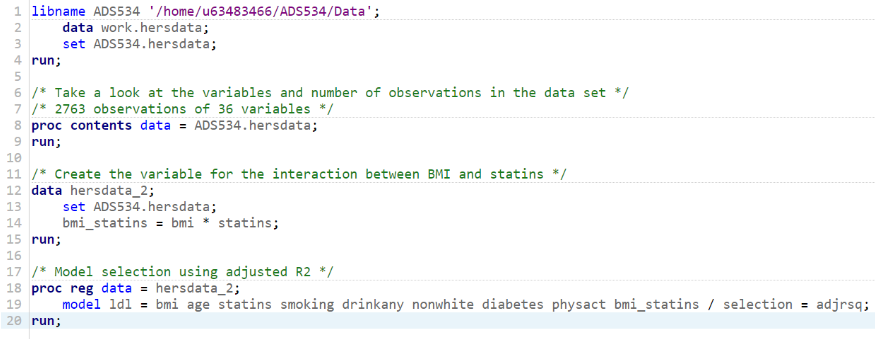
1. We will use HERS data set. We consider the following variable selection procedures.
   - Adjusted R^2
   - AIC
   - Stepwise variable selection

   The potential predictors include: BMI, Age, Statins, Smoking, Drinkany, nonwhite, diabetes, physical activities, and interaction between BMI and Statins.

   1.1. Model selection using adjusted R^2

   What is the final model based on adjusted R^2 criterion?

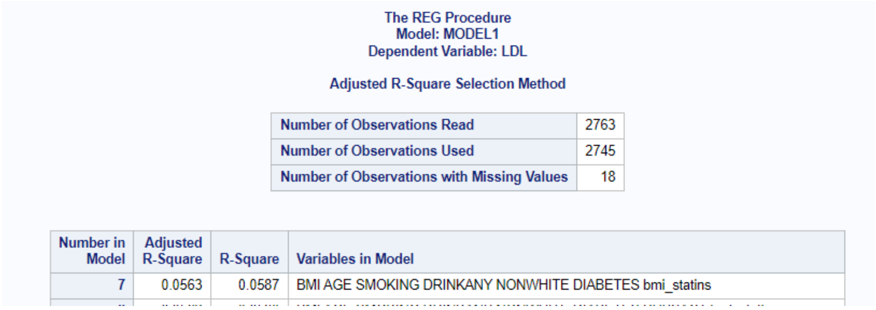


```
libname ADS534 '/home/u63483466/ADS534/Data';
    data work.hersdata;
    set ADS534.hersdata;
run;

/* Take a look at the variables and number of observations in the data set */
/* 2763 observations of 36 variables */
proc contents data = ADS534.hersdata;
run;

/* Create the variable for the interaction between BMI and statins */
data hersdata_2;
    set ADS534.hersdata;
    bmi_statins = bmi * statins;
run;

/* Model selection using adjusted R2 */
proc reg data = hersdata_2;
    model ldl = bmi age statins smoking drinkany nonwhite diabetes physact bmi_statins / selection = adjrsq;
run;
```



The REG Procedure
Model: MODEL1
Dependent Variable: LDL

Adjusted R-Square Selection Method

Number of Observations Read	2763
Number of Observations Used	2745
Number of Observations with Missing Values	18

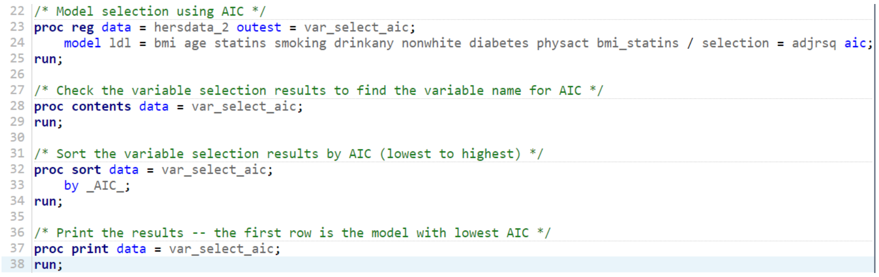
Number in Model	Adjusted R-Square	R-Square	Variables in Model
7	0.0563	0.0587	BMI AGE SMOKING DRINKANY NONWHITE DIABETES bmi_statins

   The final model based on adjusted R^2 criterion contains the variables bmi, age, smoking, drinkany, nonwhite, diabetes, and bmi_statins.

   1.2. Model selection using AIC

   Model selection using AIC is not that straightforward in SAS, we need to output variable selection results based on AIC. Then we sort the output dataset by AIC (lowest to highest). The first row of the dataset is the model with the lowest AIC.

   What is the final model based on AIC criterion?

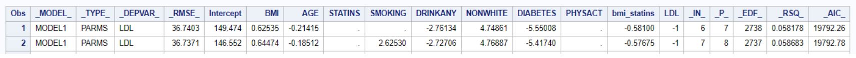


```
/* Model selection using AIC */
proc reg data = hersdata_2 outest = var_select_aic;
    model ldl = bmi age statins smoking drinkany nonwhite diabetes physact bmi_statins / selection = adjrsq aic;
run;

/* Check the variable selection results to find the variable name for AIC */
proc contents data = var_select_aic;
run;

/* Sort the variable selection results by AIC (lowest to highest) */
proc sort data = var_select_aic;
    by _AIC_;
run;

/* Print the results -- the first row is the model with lowest AIC */
proc print data = var_select_aic;
run;
```

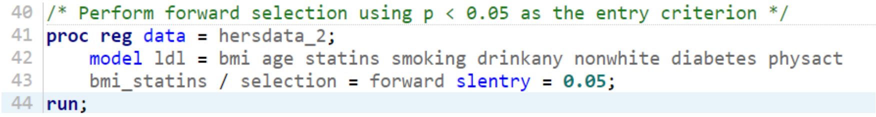


Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	BMI	AGE	STATINS	SMOKING	DRINKANY	NONWHITE	DIABETES	PHYSACT	bmi_statins	LDL	_IN_	_P_	_EDF_	_RSQ_	_AIC_
1	MODEL1	PARMS	LDL	36.7403	149.474	0.62535	-0.21415	.	.	-2.76134	4.74861	-5.55008	.	-0.58100	-1	6	7	2738	0.058178	19792.26
2	MODEL1	PARMS	LDL	36.7371	146.552	0.64474	-0.18512	.	2.62530	-2.72706	4.76887	-5.41740	.	-0.57675	-1	7	8	2737	0.058683	19792.78

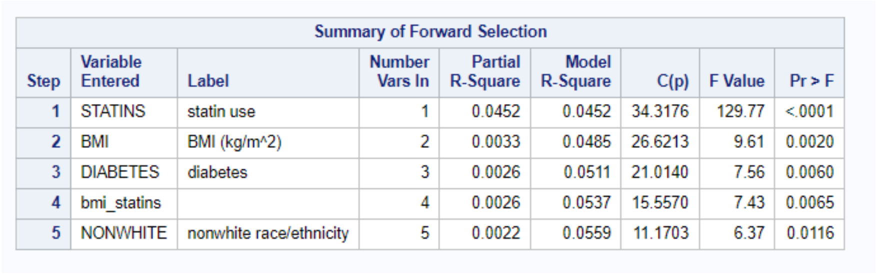
   Based on AIC criterion, the final model contains the variables bmi, age, drinkany, nonwhite, diabetes, and bmi_statins.

   1.3. Forward selection

Using p-value < 0.05 as the entry criterion. The p-value here is based on a partial F-test for a single variable. Look at the details of SAS output: which variables are selected in the first step, in the second step ...?



```
/* Perform forward selection using p < 0.05 as the entry criterion */
proc reg data = hersdata_2;
    model ldl = bmi age statins smoking drinkany nonwhite diabetes physact
    bmi_statins / selection = forward slentry = 0.05;
run;
```



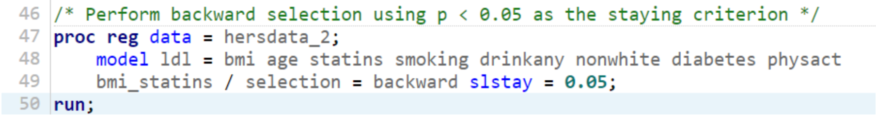
Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	STATINS	statin use	1	0.0452	0.0452	34.3176	129.77	<.0001
2	BMI	BMI (kg/m^2)	2	0.0033	0.0485	26.6213	9.61	0.0020
3	DIABETES	diabetes	3	0.0026	0.0511	21.0140	7.56	0.0060
4	bmi_statins		4	0.0026	0.0537	15.5570	7.43	0.0065
5	NONWHITE	nonwhite race/ethnicity	5	0.0022	0.0559	11.1703	6.37	0.0116

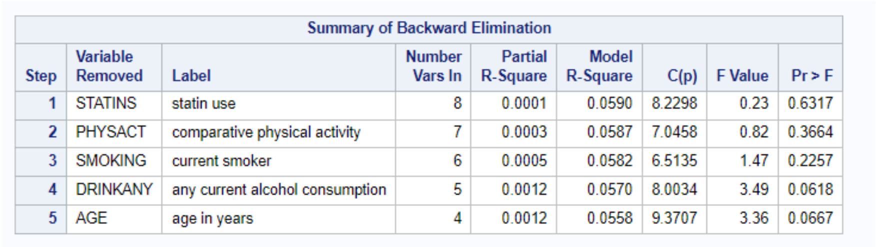
The final variables in the model based on forward selection are statins, bmi, diabetes, bmi_statins, and nonwhite. Step 1 selected statins, step 2 selected bmi, step 3 selected bmi, step 4 selected bmi_statins, and step 5 selected nonwhite.

1.4. Backward selection

Using p-value < 0.05 as the staying criterion. The p-value here is based on a partial F-test for a single variable. Look at the details of SAS output: which variables are kicked out in the first step, in the second step ...?



```
/* Perform backward selection using p < 0.05 as the staying criterion */
proc reg data = hersdata_2;
    model ldl = bmi age statins smoking drinkany nonwhite diabetes physact
    bmi_statins / selection = backward slstay = 0.05;
run;
```



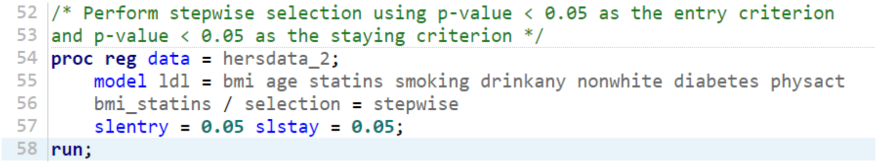
Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	STATINS	statin use	8	0.0001	0.0590	8.2298	0.23	0.6317
2	PHYSACT	comparative physical activity	7	0.0003	0.0587	7.0458	0.82	0.3664
3	SMOKING	current smoker	6	0.0005	0.0582	6.5135	1.47	0.2257
4	DRINKANY	any current alcohol consumption	5	0.0012	0.0570	8.0034	3.49	0.0618
5	AGE	age in years	4	0.0012	0.0558	9.3707	3.36	0.0667

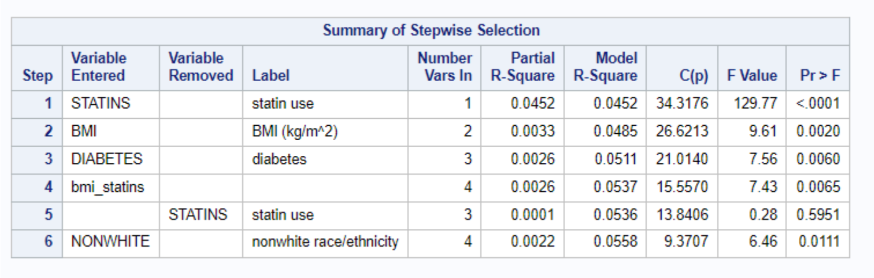
The final variables in the model based on forward selection are bmi, diabetes, bmi_statins, and nonwhite. Step 1 removed statins, step 2 removed physact, step 3 removed smoking, step 4 removed drinkany, and step 5 removed age.

1.5. Stepwise model selection

Using the stepwise selection procedure with p-value < 0.05 as the entry criterion and p-value < 0.05 as the staying criterion, what is the final model selected?



```
/* Perform stepwise selection using p-value < 0.05 as the entry criterion
and p-value < 0.05 as the staying criterion */
proc reg data = hersdata_2;
    model ldl = bmi age statins smoking drinkany nonwhite diabetes physact
    bmi_statins / selection = stepwise
    slentry = 0.05 slstay = 0.05;
run;
```



Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	STATINS		statin use	1	0.0452	0.0452	34.3176	129.77	<.0001
2	BMI		BMI (kg/m^2)	2	0.0033	0.0485	26.6213	9.61	0.0020
3	DIABETES		diabetes	3	0.0026	0.0511	21.0140	7.56	0.0060
4	bmi_statins			4	0.0026	0.0537	15.5570	7.43	0.0065
5		STATINS	statin use	3	0.0001	0.0536	13.8406	0.28	0.5951
6	NONWHITE		nonwhite race/ethnicity	4	0.0022	0.0558	9.3707	6.46	0.0111

The final model contains statins, bmi, diabetes, bmi_statins, and nonwhite. Step 1 selected statins, step 2 selected bmi, step 3 selected diabetes, step 4 selected bmi_statins, step 5 removed statins, and step 6 selected nonwhite.