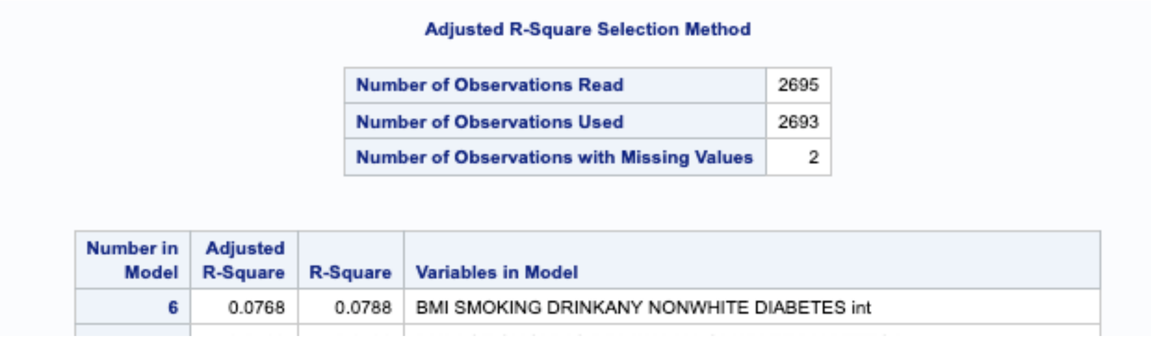
Patrick Fitzgerald
ADS 534 Statistical Modeling
Lab # 5

1.1 Model selection using adjusted R^2

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



Adjusted R-Square Selection Method

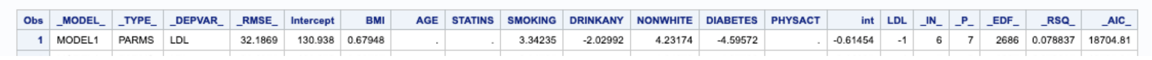
Number of Observations Read	2695
Number of Observations Used	2693
Number of Observations with Missing Values	2

Number in Model	Adjusted R-Square	R-Square	Variables in Model
6	0.0768	0.0788	BMI SMOKING DRINKANY NONWHITE DIABETES int

After removing outliers based on the observations with Studentized Residuals > 2.5, we see that the model chosen based on adjusted R^2 is BMI, Smoking, Drinkany, Nonwhite, Diabetes, and the Interaction term (BMI*Statins). (R^2 = 0.0768).

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?

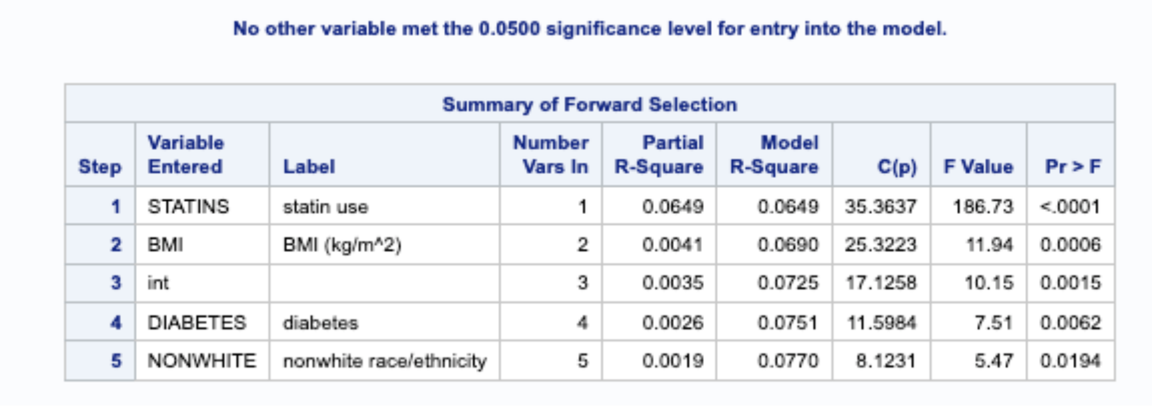


Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	BMI	AGE	STATINS	SMOKING	DRINKANY	NONWHITE	DIABETES	PHYSACT	int	LDL	_IN_	_P_	_EDF_	_RSQ_	_AIC_
1	MODEL1	PARMS	LDL	32.1869	130.938	0.67948	.	.	3.34235	-2.02992	4.23174	-4.59572	.	-0.61454	-1	6	7	2686	0.078837	18704.81

We see that after removing outliers the final model that minimizes AIC is BMI, Smoking, Nonwhite, Diabetes, and the Interaction term (BMI*Statins). (AIC = 18704.81)

1.3 Forward selection

Using p-value < 0.05 as the entry criterion. The p-value here is based on 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 ...?



No other variable met the 0.0500 significance level for entry into the model.

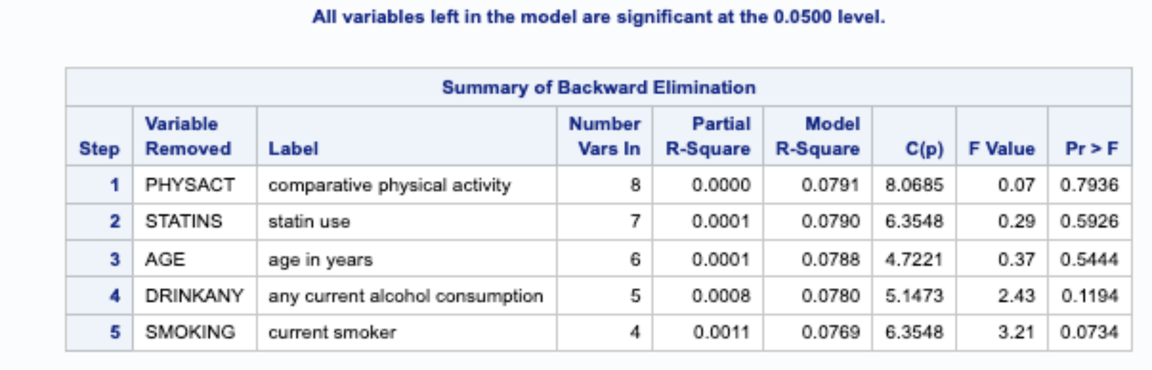
Summary of Forward Selection

Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	STATINS	statin use	1	0.0649	0.0649	35.3637	186.73	<.0001
2	BMI	BMI (kg/m^2)	2	0.0041	0.0690	25.3223	11.94	0.0006
3	int		3	0.0035	0.0725	17.1258	10.15	0.0015
4	DIABETES	diabetes	4	0.0026	0.0751	11.5984	7.51	0.0062
5	NONWHITE	nonwhite race/ethnicity	5	0.0019	0.0770	8.1231	5.47	0.0194

Forward begins with only the intercept term and adds variables to the model based on their univariate (t-test) contribution of each variable. Once a variable is added it cannot then be removed. The final model includes Statins, BMI, the Interaction term (BMI*Statins), Diabetes, and Nonwhite, which would have been added in that order. This is consistent with what we've seen so far from our various tests with some notable differences, however.

1.4 Backward selection

Using p-value < 0.05 as the staying criterion. The p-value here is based on 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 ...?



All variables left in the model are significant at the 0.0500 level.

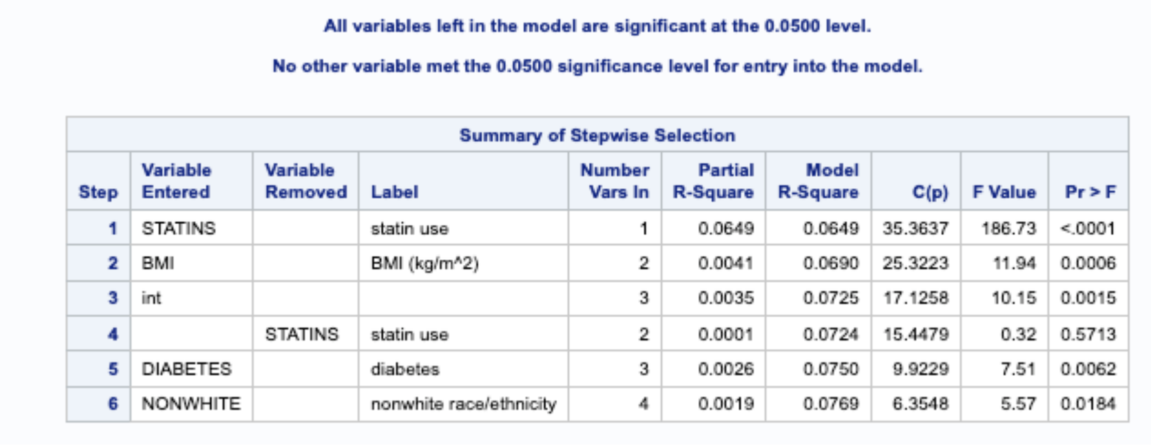
Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	PHYSACT	comparative physical activity	8	0.0000	0.0791	8.0685	0.07	0.7936
2	STATINS	statin use	7	0.0001	0.0790	6.3548	0.29	0.5926
3	AGE	age in years	6	0.0001	0.0788	4.7221	0.37	0.5444
4	DRINKANY	any current alcohol consumption	5	0.0008	0.0780	5.1473	2.43	0.1194
5	SMOKING	current smoker	4	0.0011	0.0769	6.3548	3.21	0.0734

In Backwards Elimination we start with all the variables in the model and remove them based on their univarite (t-test) contribution to the model's effectiveness. Similar to Forward Elimination where once a variable was added to the model it could not be removed, in Backwards Elimination once a variable is removed from a model it cannot be re-added. The final model includes Physact, Statins, Age, Drinkany, and Smoking. This has a few similar variables that we've seen before with some key differences.

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?



All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

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	0.0649	0.0649	35.3637	186.73	<.0001
2	BMI		BMI (kg/m^2)	2	0.0041	0.0690	25.3223	11.94	0.0006
3	int			3	0.0035	0.0725	17.1258	10.15	0.0015
4		STATINS	statin use	2	0.0001	0.0724	15.4479	0.32	0.5713
5	DIABETES		diabetes	3	0.0026	0.0750	9.9229	7.51	0.0062
6	NONWHITE		nonwhite race/ethnicity	4	0.0019	0.0769	6.3548	5.57	0.0184

Stepwise Selection is a combination of Forward and Backwards Elimination in that it starts with only the intercept term and the univariate (t-test) contributions calculated at each step determine whether a variable is added (or dropped) to (from) the model. Unlike either Forward or Backwards Elimination, in Stepwise Selection variables can be added and removed in real-time as the model updates to determine effectiveness based on the combination of variables currently comprising the model. The final model included Statins, BMI, the Interaction term (BMI*Statins), then Statins was removed, Diabetes, and Nonwhite, which is similar to the model produced using AIC only excluding Smoking.

Without additional substantive knowledge on these variables it is difficult to come to definitive conclusions despite the relative precision of the tests performed.