Bo Zhang 10411943 bzhang43@stevens.edu

1.(1)

forward method:

This is part of the results of forward method for the purpose of predicting gender ratio:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	100108	50054	66.26	<.0001
Error	787	594553	755.46702		
Corrected Total	789	694660			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.35247	16.02981	16.27076	0.02	0.8834
PC_18_65	1.58743	0.21990	39369	52.11	<.0001
PCT_O65	-0.66094	0.24847	5345.43257	7.08	0.0080

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	PC_18_65	1	0.1364	0.1364	6.3232	124.48	<.0001
2	PCT_O65	2	0.0077	0.1441	1.2632	7.08	0.0080

During the forward selection procedure, the p values of variables PC_18_65 and PCT_O65 meet the 0.05000 significance level, so they were included in the model.

Hence Model **A** is:

$y = 2.35247 + 1.58743 * (PC\_18\_65) - 0.66094 * (PCT\_O65)$

backward method:

The tables below are part of the result of backward method for the purpose of predicting gender ratio:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	100106	50053	66.25	<.0001
Error	787	594554	755.46851		
Corrected Total	789	694660			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	161.08957	6.91759	409677	542.28	<.0001
PCT_U18	-1.58719	0.21987	39368	52.11	<.0001
PCT_O65	-2.24835	0.20764	88579	117.25	<.0001

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	PC_18_65	3	0.0000	0.1444	3.0001	0.00	0.9903
2	TOT_POP	2	0.0003	0.1441	1.2647	0.26	0.6069

During the backward elimination procedure, the p values of PC__18_65 and TOT_POP is more than 0.05 significance level, so they were dropped from the model.

Hence Model **B** is:

y =161.08957-1.58719*(*PCT_U18*)-2.24835 *(*PCT_O65*)

stepwise method:

This is part of the results of stepwise method for the purpose of predicting gender ratio:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	100108	50054	66.26	<.0001
Error	787	594553	755.46702		
Corrected Total	789	694660			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.35247	16.02981	16.27076	0.02	0.8834
PC_18_65	1.58743	0.21990	39369	52.11	<.0001
PCT_065	-0.66094	0.24847	5345.43257	7.08	0.0080

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	PC_18_65		1	0.1364	0.1364	6.3232	124.48	<.0001
2	PCT_065		2	0.0077	0.1441	1.2632	7.08	0.0080

Hence Model **C** is:

*y = 2.35247 + 1.58743**(*PC_18_65*) -0.66094*(*PCT_O65*)

Comparing and contrasting the three models:

- During the forward selection procedure, the p values of variables PC_18_65 and PCT_O65 meet the 0.05000 significance level, so they were included in the model. The forward selection procedure did not include the following variables: PCT_U18 and TOT_POP.
- During the backward selection procedure, the p values of PC_18_65 and TOT_POP didn't meet the 0.05000 significance level for entry into the model, so they were dropped from the model one by one, so finally the procedure included PCT_U18 and PCT_O65.
- The stepwise selection is a modification of the forward selection procedure, each variable that had been entered remained significant when the other variables were also entered. Therefore, the models' results were the same as for the forward selection.

1.(2)

This is the best 2-variable model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	100108	50054	66.26	<.0001
Error	787	594553	755.46702		
Corrected Total	789	694660			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.35247	16.02981	16.27076	0.02	0.8834
PC_18_65	1.58743	0.21990	39369	52.11	<.0001
PCT_065	-0.66094	0.24847	5345.43257	7.08	0.0080

This is the best 3-variable model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	100307	33436	44.22	<.0001
Error	786	594353	756.17482		
Corrected Total	789	694660			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	161.19702	6.92397	409850	542.00	<.0001
TOT_POP	-0.00000193	0.00000375	200.30567	0.26	0.6069
PCT_U18	-1.58914	0.22001	39453	52.17	<.0001
PCT_065	-2.25012	0.20776	88694	117.29	<.0001

During the best subset procedure:

- When variable PC_18_65 in the model, and its p value is less than 0.0001. So the best 1-variable model is :
  y =-29.25510+1.94596*(PC_18_65)
- When variable PCT_O65 entered the model, there were PCT_O65 and PC_18_65 in the model, both of the variables met the 0.05 significance level. So the best 2-variable model is:
  y = 2.35247 +1.58743*(PC_18_65) -0.66094*(PCT_O65), this is the same as model **A** (**C**)
- When variable TOT_POP entered the model, there were just two variables PC_18_65 and PCT_O65 met 0.05 significance level. And when variable PC_18_65 removed and PCT_U18 entered the model, there were PCT_U18 and PCT_O65 met the 0.05 significance level. So the best 3-variable model is:
  y = 161.19702-1.58914*(PCT_U18)-2.25012*(PCT_O65), this is the same as model **B**
- When variable PC_18_65 reentered the model, all the p values of variables were high and none of them were significant.

Finally, comparing model **A**(**C**)and model **B**, the p values of model B were all less than 0.0001, however, the p value of in model **A**(**C**) is 0.0080. So the best model is model **B**:
That is: y= 161.19702-1.58914*(PCT_U18)-2.25012*(PCT_O65).

2.(1)

forward method:

This is part of results, for the purpose of predicting calories:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	27975	3108.36039	237.55	<.0001
Error	67	876.70453	13.08514		
Corrected Total	76	28852			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	211.37235	11.34221	4544.42763	347.30	<.0001
sugars	-2.45935	0.35341	633.66162	48.43	<.0001
fiber	9.23453	0.64207	2706.72934	206.86	<.0001
shelf	-0.68025	0.59842	16.90806	1.29	0.2597
sodium	-0.19917	0.01306	3044.69563	232.68	<.0001
fat	-5.57178	0.97166	430.26859	32.88	<.0001
protein	11.70063	0.61019	4811.41603	367.70	<.0001
carbo	4.34525	0.16177	9441.24130	721.52	<.0001
vitamins	-0.17958	0.02540	653.96368	49.98	<.0001
rating	-3.66875	0.21485	3815.52903	291.59	<.0001

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	rating	1	0.4752	0.4752	1084.07	67.92	<.0001
2	protein	2	0.1516	0.6268	751.887	30.05	<.0001
3	carbo	3	0.1305	0.7573	466.127	39.25	<.0001
4	sodium	4	0.0520	0.8094	353.360	19.66	<.0001
5	fiber	5	0.1232	0.9325	83.7690	129.62	<.0001
6	vitamins	6	0.0136	0.9461	55.8326	17.63	<.0001
7	sugars	7	0.0061	0.9522	44.3536	8.83	0.0041

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
8	fat	8	0.0168	0.9690	9.2922	36.90	<.0001
9	shelf	9	0.0006	0.9696	10.0000	1.29	0.2597

From the table, the p value of shelf is 0.2597, it is more than 0.05, so the model should not include variable shelf.
So Model **A** is:
y=211.37235-2.45935(sugars)*+9.23453*(fiber)-0.19917*(*sodium*)-5.571 78*(fat)+11.70063*(*protein*)+4.34525*(*carbo*)-0.17958*(*vitamins*)-3.668 75*(rating)

backward method:
The tables below are part of the results, backward method for the purpose of predicting calories:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	27958	3494.79193	265.94	<.0001
Error	68	893.61260	13.14136		
Corrected Total	76	28852			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	211.01002	11.36206	4532.44075	344.90	<.0001
fat	-5.79415	0.95380	484.95405	36.90	<.0001
fiber	9.17155	0.64105	2689.96509	204.69	<.0001
sodium	-0.19823	0.01306	3028.16851	230.43	<.0001
protein	11.77696	0.60778	4934.14640	375.47	<.0001
carbo	4.33234	0.16171	9431.72084	717.71	<.0001
vitamins	-0.18947	0.02392	824.43701	62.74	<.0001
sugars	-2.48406	0.35350	648.91533	49.38	<.0001

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
rating	-3.68010	0.21508	3847.50447	292.78	<.0001

During the backward selection procedure, the p values of shelf did not meet the 0.05000 significance level for entry into the model, so they were dropped from the model one by one.

So Model **B** is:

y=211.01002-5.79415(fat)+9.17155*(fiber)-0.19823*(sodium)* +11.77696*(*protein*)+4.332348*(*cardo*)-0.18947*(*vitamins*)-2.48406*(sugars)-3.68010*(*rating*)

stepwise method:

The tables below are part of the results, stepwise method for the purpose of predicting calories:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	27958	3494.79193	265.94	<.0001
Error	68	893.61260	13.14136		
Corrected Total	76	28852			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	211.01002	11.36206	4532.44075	344.90	<.0001
fat	-5.79415	0.95380	484.95405	36.90	<.0001
fiber	9.17155	0.64105	2689.96509	204.69	<.0001
sodium	-0.19823	0.01306	3028.16851	230.43	<.0001
protein	11.77696	0.60778	4934.14640	375.47	<.0001
carbo	4.33234	0.16171	9431.72084	717.71	<.0001
vitamins	-0.18947	0.02392	824.43701	62.74	<.0001
rating	-3.68010	0.21508	3847.50447	292.78	<.0001
sugars	-2.48406	0.35350	648.91533	49.38	<.0001

So Model **C** is:
y=211.01002-5.79415(fat)+9.17155*(fiber)-0.19823*(sodium)*
+11.77696*(*protein*)+4.332348*(*cardo*)-0.18947*(*vitamins*)-2.48406*(sugars)-3.68010*(*rating*)
And it is the same with Model **B**.

best subset method:
This is the best 8-variable model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	27958	3494.79193	265.94	<.0001
Error	68	893.61260	13.14136		
Corrected Total	76	28852			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	211.01002	11.36206	4532.44075	344.90	<.0001
fat	-5.79415	0.95380	484.95405	36.90	<.0001
fiber	9.17155	0.64105	2689.96509	204.69	<.0001
sodium	-0.19823	0.01306	3028.16851	230.43	<.0001
protein	11.77696	0.60778	4934.14640	375.47	<.0001
carbo	4.33234	0.16171	9431.72084	717.71	<.0001
vitamins	-0.18947	0.02392	824.43701	62.74	<.0001
rating	-3.68010	0.21508	3847.50447	292.78	<.0001
sugars	-2.48406	0.35350	648.91533	49.38	<.0001

This is the best 9-variable model:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	27975	3108.36039	237.55	<.0001

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Error	67	876.70453	13.08514		
Corrected Total	76	28852			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	211.37235	11.34221	4544.42763	347.30	<.0001
fat	-5.57178	0.97166	430.26859	32.88	<.0001
fiber	9.23453	0.64207	2706.72934	206.86	<.0001
shelf	-0.68025	0.59842	16.90806	1.29	0.2597
sodium	-0.19917	0.01306	3044.69563	232.68	<.0001
protein	11.70063	0.61019	4811.41603	367.70	<.0001
carbo	4.34525	0.16177	9441.24130	721.52	<.0001
vitamins	-0.17958	0.02540	653.96368	49.98	<.0001
rating	-3.66875	0.21485	3815.52903	291.59	<.0001
sugars	-2.45935	0.35341	633.66162	48.43	<.0001

From the above procedure, the p values of best 8-variables were all less than 0.0001, whereas, the p value of shelf in the best 9-variable model is 0.0005. And they were separately model **A** and model **B**(**C**)

2.(2)

Comparing and contrasting the three models:

- During the forward selection procedure, the p values of variables fiber, sodium, protein, cargo, vitamins,sugars and rating met the 0.05000 significance level, so they were included in the model.
- During the backward elimination procedure, the variable shelf were dropped form the model, due to the p value is not significant.
- Although the stepwise selection is a modification of the forward selection procedure, but it has the same model with back elimination procedure.

As a result, comparing to model **B**(**C**), model **A** has higher F vlaue. Therefore, the final model is model **A**,

y=211.37235-2.45935(sugars)*+9.23453*(fiber)-0.19917*(*sodium*)-5.571 78*(fat)+11.70063*(*protein*)+4.34525*(*carbo*)-0.17958*(*vitamins*)-3.668 75*(rating)