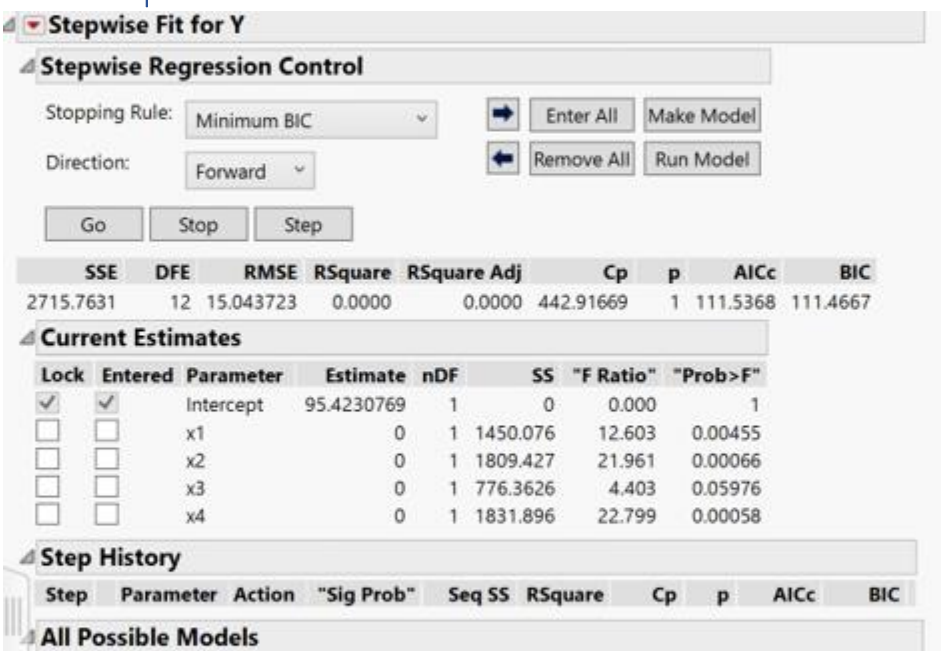
Exercise 2

a. We can infer from table 7.6 that subset size 3 has better R^2 adjusted and AIC. Whereas, the other AIC_C and BIC conclude the model with subset size 2 is better. Out of these two, you cannot tell which is better with just this data. More analysis is required.
b. For forward selection the best model would be subset 3 – due to the AIC, BIC values
c. For backward selection, AIC gives 3 subset model, BIC gives 2 subset model
d. The methodologies of selection are different. In a – it was the best of all 2^4 subsets. In b,c – once a predictor is added or dropped, it is constrained. Hence, all the best models are different.
e. Two subset model (Y ~ X1 + X2) is the best. It can be inferred from the summary statistics – R^2 adjusted and significance of predictors or p-value. One would be inclined to go with Four subset size model. But, due to the collinearity between X1 and X4, as a result high VIF, we reject this model in spite of having a high R^2 adjusted.

JMP outputs



The JMP results are more accurate than R. It is due to the difference in formulas of penalty terms for AIC and BIC. These are inbuilt in JMP, whereas it is not in R.

Model	Number	RSquare	RMSE	AICc	BIC
x4	1	0.6745	8.9639	100.411	99.4389
x1,x2	2	0.9787	2.4063	69.3124	66.5722
x1,x2,x4	3	0.9823	2.3087	72.4377	66.6910
x1,x2,x3,x4	4	0.9824	2.4460	79.8367	69.2264

Python

```
import pandas as pd
import numpy  as np
```

```
# sklearn packages
from sklearn.linear_model import LinearRegression
# Statsmodel
import statsmodels.api as sm
from AdvancedAnalytics import ReplaceImputeEncode
from AdvancedAnalytics import linreg
df = pd.read_csv("Haldcement.csv",delimiter="\t")
X   = np.asarray(df.drop('Y', axis=1))
X1  = np.asarray(df.drop(['Y', 'x2', 'x3','x4'], axis=1))
X2  = np.asarray(df.drop(['Y', 'x1', 'x3','x4'], axis=1))
X3  = np.asarray(df.drop(['Y', 'x1', 'x2','x4'], axis=1))
X4=np.asarray(df.drop(['Y', 'x1', 'x2','x3'], axis=1))
X12 = np.asarray(df.drop(['Y', 'x3','x4'], axis=1))
X13 = np.asarray(df.drop(['Y', 'x2','x4'], axis=1))
X23 = np.asarray(df.drop(['Y', 'x1','x4'], axis=1))
X14=np.asarray(df.drop(['Y', 'x2','x3'], axis=1))
X24=np.asarray(df.drop(['Y', 'x1','x3'], axis=1))
X34=np.asarray(df.drop(['Y', 'x1','x2'], axis=1))
X123=np.asarray(df.drop(['Y','x4'], axis=1))
X234=np.asarray(df.drop(['Y','x1'], axis=1))
X134=np.asarray(df.drop(['Y','x2'], axis=1))
X124=np.asarray(df.drop(['Y','x3'], axis=1))
y   = np.asarray(df['Y'])
print(df)
col = ['X1', 'X2', 'X3','X4']
lr = LinearRegression()
lr.fit(X, y)
print("\nLinear Regression")
linreg.display_coef(lr, X, y, col)
linreg.display_metrics(lr, X, y)

print("\nStats Model Fit:\n")
Xc = sm.add_constant(X)
ols_model = sm.OLS(y, Xc)
results   = ols_model.fit()
print(results.summary())
lr.fit(X1, y)
print("\nLinear Regression for ", col[0])
linreg.display_coef(lr, X1, y, [col[0]])
linreg.display_metrics(lr, X1, y)
lr.fit(X2, y)
print("\nStats Model Fit:\n")
Xc = sm.add_constant(X1)
ols_model = sm.OLS(y, Xc)
results   = ols_model.fit()
print(results.summary())

print("\nLinear Regression for ", col[1])
linreg.display_coef(lr, X2, y, [col[1]])
linreg.display_metrics(lr, X2, y)
```

```
lr.fit(X3, y)
print("\nStats Model Fit:\n")
Xc = sm.add_constant(X2)
ols_model = sm.OLS(y, Xc)
results   = ols_model.fit()
print(results.summary())

print("\nLinear Regression for ", col[2])
linreg.display_coef(lr, X3, y, [col[2]])
linreg.display_metrics(lr, X3, y)
lr.fit(X4, y)
print("\nStats Model Fit:\n")
Xc = sm.add_constant(X3)
ols_model = sm.OLS(y, Xc)
results   = ols_model.fit()
print(results.summary())
print("\nLinear Regression for ", col[3])
linreg.display_coef(lr, X4, y, [col[3]])
linreg.display_metrics(lr, X4, y)

print("\nStats Model Fit:\n")
Xc = sm.add_constant(X4)
ols_model = sm.OLS(y, Xc)
results   = ols_model.fit()
print(results.summary())
```

```
Linear Regression

Coefficients

Coefficients
Intercept..        62.4054
X1.........         1.5511
X2.........         0.5102
X3.........         0.1019
X4.........        -0.1441

Model Metrics
Observations...........               13
Coefficients...........                5
DF Error...............                8
R-Squared..............           0.9824
Mean Absolute Error....           1.5871
Median Absolute Error..           1.5112
Avg Squared Error......           3.6818
Square Root ASE........           1.9188

Stats Model Fit:

                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.982
Model:                            OLS   Adj. R-squared:                  0.974
Method:                 Least Squares   F-statistic:                     111.5
Date:                Wed, 03 Apr 2019   Prob (F-statistic):           4.76e-07
Time:                        09:00:06   Log-Likelihood:                -26.918
No. Observations:                  13   AIC:                             63.84
Df Residuals:                       8   BIC:                             66.66
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         62.4054     70.071      0.891      0.399     -99.179     223.989
x1             1.5511      0.745      2.083      0.071      -0.166       3.269
x2             0.5102      0.724      0.705      0.501      -1.159       2.179

x3             0.1019      0.755      0.135      0.896      -1.638       1.842
x4            -0.1441      0.709     -0.203      0.844      -1.779       1.491
==============================================================================
Omnibus:                        0.165   Durbin-Watson:                   2.053
Prob(Omnibus):                  0.921   Jarque-Bera (JB):                0.320
Skew:                           0.201   Prob(JB):                        0.852
Kurtosis:                       2.345   Cond. No.                     6.06e+03
==============================================================================
```

Exercise 3

a.

Stepwise Fit for Log prize money

Stepwise Regression Control

Stopping Rule: Minimum BIC

Direction: Forward

Enter All | Make Model | Remove All | Run Model | Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
187.35515	195	0.9802018	0.0000	0.0000	231.05923	1	551.4448	557.9388

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	10.3780793	1	0	0.000	1
		DrivingAccuracy	0	1	6.183665	6.622	0.01082
		GIR	0	1	47.76003	66.374	4.5e-14
		PuttingAverage	0	1	34.65997	44.036	3.1e-10
		BirdieConversion	0	1	40.92997	54.228	5e-12
		SandSaves	0	1	10.92588	12.014	0.00065
		Scrambling	0	1	25.26049	30.233	1.2e-7
		PuttsPerRound	0	1	6.294789	6.745	0.01012

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC

All Possible Models

Ordered up to best 1 models up to 7 terms per model.

Model	Number	RSquare	RMSE	AICc	BIC
GIR	1	0.2549	0.8483	495.833	505.542
GIR,PuttsPerRound	2	0.4910	0.7029	423.213	436.116
GIR,BirdieConversion,Scrambling	3	0.5453	0.6661	403.229	419.304
GIR,BirdieConversion,SandSaves,Scrambling	4	0.5522	0.6628	402.377	421.602
GIR,BirdieConversion,SandSaves,Scrambling,PuttsPerRound	5	0.5575	0.6606	402.178	424.529
DrivingAccuracy,GIR,BirdieConversion,SandSaves,Scrambling,PuttsPerRound	6	0.5577	0.6622	404.263	429.718
DrivingAccuracy,GIR,PuttingAverage,BirdieConversion,SandSaves,Scrambling,PuttsPerRound	7	0.5577	0.6639	406.456	434.991

The 5th model from the top is the best. It is where we see a transition in RMSE and min AICc, R square. If you look at min BIC – has 3 predictors, min AIC – 5 predictictors

b. Based on backward selection, and using BIC as the criteria, we would go with the 3rd model

Stepwise Fit for Log prize money

Stepwise Regression Control

Stopping Rule: Minimum BIC

Direction: Backward

Enter All | Make Model | Remove All | Run Model | Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
85.191057	192	0.6661107	0.5453	0.5382	5.2759549	4	403.2289	419.3037

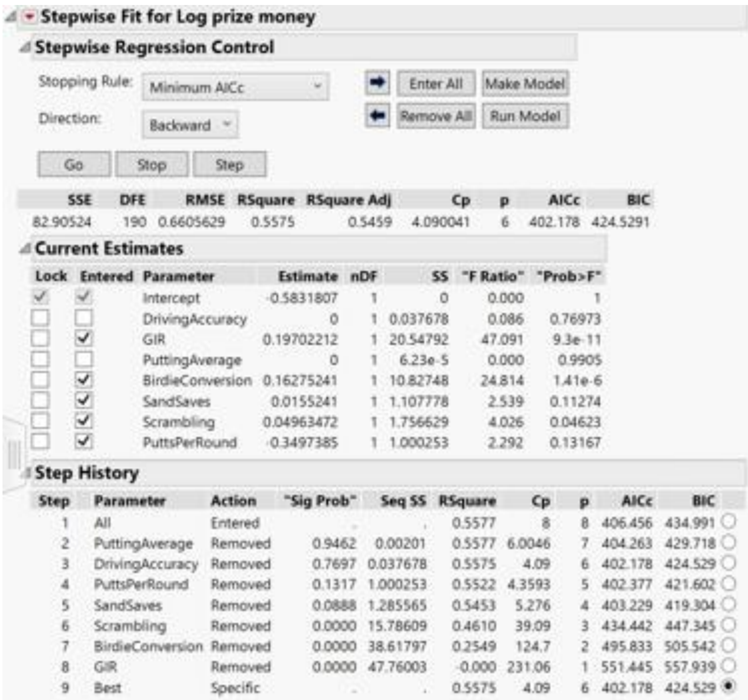
Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	-11.083136	1	0	0.000	1
		DrivingAccuracy	0	1	4.424e-5	0.000	0.99206
	✓	GIR	0.15658127	1	34.05712	76.756	1e-15
		PuttingAverage	0	1	0.877209	1.987	0.16026
	✓	BirdieConversion	0.20624558	1	40.30825	90.845	7.1e-18
		SandSaves	0	1	1.285565	2.926	0.08876
	✓	Scrambling	0.09177979	1	15.78609	35.578	1.16e-8
		PuttsPerRound	0	1	1.178039	2.678	0.10338

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	All	Entered	.	.	0.5577	8	8	406.456	434.991
2	PuttingAverage	Removed	0.9462	0.00201	0.5577	6.0046	7	404.263	429.718
3	DrivingAccuracy	Removed	0.7697	0.037678	0.5575	4.09	6	402.178	424.529
4	PuttsPerRound	Removed	0.1317	1.000253	0.5522	4.3593	5	402.377	421.602
5	SandSaves	Removed	0.0888	1.285565	0.5453	5.276	4	403.229	419.304
6	Scrambling	Removed	0.0000	15.78609	0.4610	39.09	3	434.442	447.345
7	BirdieConversion	Removed	0.0000	38.61797	0.2549	124.7	2	495.833	505.542
8	GIR	Removed	0.0000	47.76003	-0.000	231.06	1	551.445	557.939
9	Best	Specific	.	.	0.5453	5.276	4	403.229	419.304

Based on AIC, the best model is 5th model – 5 predictors



Stepwise Fit for Log prize money

Stepwise Regression Control

Stopping Rule: Minimum AICc

Direction: Backward

Enter All | Make Model | Remove All | Run Model

Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
82.90524	190	0.6605629	0.5575	0.5459	4.090041	6	402.178	424.5291

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	-0.5831807	1	0	0.000	1
☐	☐	DrivingAccuracy	0	1	0.037678	0.086	0.76973
☐	✓	GIR	0.19702212	1	20.54792	47.091	9.3e-11
☐	☐	PuttingAverage	0	1	6.23e-5	0.000	0.9905
☐	✓	BirdieConversion	0.16275241	1	10.82748	24.814	1.41e-6
☐	✓	SandSaves	0.0155241	1	1.107778	2.539	0.11274
☐	✓	Scrambling	0.04963472	1	1.756629	4.026	0.04623
☐	✓	PuttsPerRound	-0.3497385	1	1.000253	2.292	0.13167

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	All	Entered	.	.	0.5577	8	8	406.456	434.991	○
2	PuttingAverage	Removed	0.9462	0.00201	0.5577	6.0046	7	404.263	429.718	○
3	DrivingAccuracy	Removed	0.7697	0.037678	0.5575	4.09	6	402.178	424.529	○
4	PuttsPerRound	Removed	0.1317	1.000253	0.5522	4.3593	5	402.377	421.602	○
5	SandSaves	Removed	0.0888	1.285565	0.5453	5.276	4	403.229	419.304	○
6	Scrambling	Removed	0.0000	15.78609	0.4610	39.09	3	434.442	447.345	○
7	BirdieConversion	Removed	0.0000	38.61797	0.2549	124.7	2	495.833	505.542	○
8	GIR	Removed	0.0000	47.76003	-0.000	231.06	1	551.445	557.939	○
9	Best	Specific	.	.	0.5575	4.09	6	402.178	424.529	●

c. Based on forward selection,

and minimum BIC – we get the model with 4 predictors.

Stepwise Fit for Log prize money

Stepwise Regression Control

Stopping Rule: Minimum BIC

Direction: Forward

Enter All | Make Model | Remove All | Run Model

Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
84.013019	191	0.6632185	0.5516	0.5422	4.6032965	5	402.6283	421.8526

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	0.39320263	1	0	0.000	1
☐	☐	DrivingAccuracy	0	1	0.099372	0.225	0.6358
☐	✓	GIR	0.19351831	1	19.94056	45.334	1.9e-10
☐	☐	PuttingAverage	0	1	0.000335	0.001	0.97807
☐	✓	BirdieConversion	0.16589366	1	11.29059	25.669	9.52e-7
☐	☐	SandSaves	0	1	1.107778	2.539	0.11274
☐	✓	Scrambling	0.06282001	1	3.168428	7.203	0.00792
☐	✓	PuttsPerRound	-0.3783971	1	1.178039	2.678	0.10338

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	All	Removed	.	.	0.0000	231.06	1	551.445	557.939	○
2	GIR	Entered	0.0000	47.76003	0.2549	124.7	2	495.833	505.542	○
3	PuttsPerRound	Entered	0.0000	44.24047	0.4910	26.334	3	423.213	436.116	○
4	BirdieConversion	Entered	0.0000	8.1732	0.5347	9.7916	4	407.756	423.83	○
5	Scrambling	Entered	0.0079	3.168428	0.5516	4.6033	5	402.628	421.853	○
6	SandSaves	Entered	0.1127	1.107778	0.5575	4.09	6	402.178	424.529	○
7	DrivingAccuracy	Entered	0.7697	0.037678	0.5577	6.0046	7	404.263	429.718	○
8	PuttingAverage	Entered	0.9462	0.00201	0.5577	8	8	406.456	434.991	○
9	Best	Specific	.	.	0.5516	4.6033	5	402.628	421.853	◉

Based on min AIC, the best model has 5 predictors

Stepwise Fit for Log prize money

Stepwise Regression Control

Stopping Rule: Minimum AICc

Direction: Forward

Enter All | Make Model | Remove All | Run Model

Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
82.90524	190	0.6605629	0.5575	0.5459	4.090041	6	402.178	424.5291

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
✓	✓	Intercept	-0.5831807	1	0	0.000	1
		DrivingAccuracy	0	1	0.037678	0.086	0.76973
	✓	GIR	0.19702212	1	20.54792	47.091	9.3e-11
		PuttingAverage	0	1	6.23e-5	0.000	0.9905
	✓	BirdieConversion	0.16275241	1	10.82748	24.814	1.41e-6
	✓	SandSaves	0.0155241	1	1.107778	2.539	0.11274
	✓	Scrambling	0.04963472	1	1.756629	4.026	0.04623
	✓	PuttsPerRound	-0.3497385	1	1.000253	2.292	0.13167

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	GIR	Entered	0.0000	47.76003	0.2549	124.7	2	495.833	505.542
2	PuttsPerRound	Entered	0.0000	44.24047	0.4910	26.334	3	423.213	436.116
3	BirdieConversion	Entered	0.0000	8.1732	0.5347	9.7916	4	407.756	423.83
4	Scrambling	Entered	0.0079	3.168428	0.5516	4.6033	5	402.628	421.853
5	SandSaves	Entered	0.1127	1.107778	0.5575	4.09	6	402.178	424.529
6	DrivingAccuracy	Entered	0.7697	0.037678	0.5577	6.0046	7	404.263	429.718
7	PuttingAverage	Entered	0.9462	0.00201	0.5577	8	8	406.456	434.991
8	Best	Specific	.	.	0.5575	4.09	6	402.178	424.529

d. Explanation for why the models chosen in (a) & (c) are not the same while those in (a) and (b) are the same.

In Question c, we use forward selection approach. This means you cannot remove the variable after entering. The predictor that enters in the 3rd step is the final model, due to minimum BIC. Whereas in Question a, the first 3 predictors that enter the model are totally different. Backward selection follows a reverse approach – you remove predictors one by one. Hence, that variable might be removed. Hence a and c are different.

E . After checking the VIF and hence multicollinearity, we can definitely go ahead with the model that has 5 predictors.

GIR,BirdieConversion,SandSaves,Scrambling,PuttsPerRound	5	0.5575	0.6606	402.178	424.529

f. The regression coefficients tell us that if the prize money would increase or decrease with the predictors chosen. It is positive for GIR, Birdie conversion, Scrambling and Sand saves. Whereas the others have negative coefficient. But, checking the VIF values reveals collinearity between Puts per round and other variables. Due to this large VIF, the coefficients are affected, thus resulting in high errors with intercepts and predictors. So, the coefficients are not reliable at this stage.

Summary of Fit

RSquare	0.557497
RSquare Adj	0.545852
Root Mean Square Error	0.660563
Mean of Response	10.37808
Observations (or Sum Wgts)	196

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	104.44991	20.8900	47.8751
Error	190	82.90524	0.4363	**Prob > F**
C. Total	195	187.35515		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	VIF
Intercept	-0.583181	7.158721	-0.08	0.9352	.
GIR	0.1970221	0.028711	6.86	<.0001*	2.7301655
BirdieConversion	0.1627524	0.032672	4.98	<.0001*	2.3226928
SandSaves	0.0155241	0.009743	1.59	0.1127	1.4410544
Scrambling	0.0496347	0.024738	2.01	0.0462*	2.7347655
PuttsPerRound	-0.349738	0.230995	-1.51	0.1317	4.6523358