Homework 3 - Linear Model Selection

Part A

1. no interactions

OLS Regression Results

Dep. Variable:	Hwt	R-squared:	0.647
Model:	OLS	Adj. R-squared:	0.642
Method:	Least Squares	F-statistic:	129.1
Date:	Tue, 17 Apr 2018	Prob (F-statistic):	1.37e-32
Time:	22:01:39	Log-Likelihood:	-257.02
No. Observations:	144	AIC:	520.0
Df Residuals:	141	BIC:	529.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.4150	0.727	-0.571	0.569	-1.853	1.023
Sex[T.M]	-0.0821	0.304	-0.270	0.788	-0.683	0.519
Bwt	4.0758	0.295	13.826	0.000	3.493	4.659

Omnibus:	4.665	Durbin-Watson:	1.581
Prob(Omnibus):	0.097	Jarque-Bera (JB):	4.245
Skew:	0.410	Prob(JB):	0.120
Kurtosis:	3.192	Cond. No.	19.6

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

2. interactions

OLS Regression Results

Dep. Variable:	Hwt	R-squared:	0.657
Model:	OLS	Adj. R-squared:	0.649
Method:	Least Squares	F-statistic:	89.24
Date:	Tue, 17 Apr 2018	Prob (F-statistic):	2.46e-32
Time:	22:01:39	Log-Likelihood:	-254.99
No. Observations:	144	AIC:	518.0
Df Residuals:	140	BIC:	529.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.9813	1.843	1.618	0.108	-0.662	6.625
Sex[T.M]	-4.1654	2.062	-2.020	0.045	-8.242	-0.089
Bwt	2.6364	0.776	3.398	0.001	1.102	4.170
Bwt:Sex[T.M]	1.6763	0.837	2.002	0.047	0.021	3.332

Omnibus:	3.872	Durbin-Watson:	1.624
Prob(Omnibus):	0.144	Jarque-Bera (JB):	3.439
Skew:	0.367	Prob(JB):	0.179
Kurtosis:	3.183	Cond. No.	91.0

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The P values are much lower in the second model including the interactions, the Adjusted R-Squared is only slightly higher, indicating that the model with interactions is only slightly better, but still better.

3. Using the model with interactions Sigma's Heart Weight is 11.945 gs. Using the model without interactions Sigma's Heart Weight is 13.443 gs.

Part B

1.

no interactions

OLS Regression Results

Dep. Variable:	Volume	R-squared:	0.948
Model:	OLS	Adj. R-squared:	0.944
Method:	Least Squares	F-statistic:	255.0
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	1.07e-18
Time:	19:04:06	Log-Likelihood:	-84.455
No. Observations:	31	AIC:	174.9
Df Residuals:	28	BIC:	179.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-57.9877	8.638	-6.713	0.000	-75.682	-40.293
Girth	4.7082	0.264	17.816	0.000	4.167	5.249
Height	0.3393	0.130	2.607	0.014	0.073	0.606

Omnibus:	0.923	Durbin-Watson:	1.266
Prob(Omnibus):	0.630	Jarque-Bera (JB):	0.950
Skew:	0.310	Prob(JB):	0.622
Kurtosis:	2.408	Cond. No.	959.

interactions

OLS Regression Results

Dep. Variable:	Volume	R-squared:	0.976
Model:	OLS	Adj. R-squared:	0.973
Method:	Least Squares	F-statistic:	359.3
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	7.29e-22
Time:	19:04:06	Log-Likelihood:	-72.735
No. Observations:	31	AIC:	153.5
Df Residuals:	27	BIC:	159.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	69.3963	23.836	2.911	0.007	20.489	118.303
Girth	-5.8558	1.921	-3.048	0.005	-9.798	-1.914
Height	-1.2971	0.310	-4.186	0.000	-1.933	-0.661
Girth:Height	0.1347	0.024	5.524	0.000	0.085	0.185

Omnibus:	2.124	Durbin-Watson:	1.993
Prob(Omnibus):	0.346	Jarque-Bera (JB):	1.463
Skew:	-0.532	Prob(JB):	0.481
Kurtosis:	3.009	Cond. No.	5.21e+04

2.

no interactions

OLS Regression Results

Dep. Variable:	Volume	R-squared:	0.896
Model:	OLS	Adj. R-squared:	0.889
Method:	Least Squares	F-statistic:	120.6
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	1.73e-14
Time:	19:04:06	Log-Likelihood:	-95.186
No. Observations:	31	AIC:	196.4
Df Residuals:	28	BIC:	200.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-234.8876	53.925	-4.356	0.000	-345.348	-124.427
np.log(Girth)	61.2687	5.058	12.114	0.000	50.909	71.629
np.log(Height)	25.0447	13.784	1.817	0.080	-3.191	53.280

Omnibus:	5.742	Durbin-Watson:	0.834
Prob(Omnibus):	0.057	Jarque-Bera (JB):	4.303
Skew:	0.882	Prob(JB):	0.116
Kurtosis:	3.468	Cond. No.	289.

interactions

OLS Regression Results

Dep. Variable:	Volume	R-squared:	0.966
Model:	OLS	Adj. R-squared:	0.962
Method:	Least Squares	F-statistic:	254.7
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	6.61e-20
Time:	19:04:06	Log-Likelihood:	-77.915
No. Observations:	31	AIC:	163.8
Df Residuals:	27	BIC:	169.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1969.8081	298.210	6.605	0.000	1357.933	2581.684
np.log(Girth)	-828.5179	119.718	-6.921	0.000	-1074.160	-582.876
np.log(Height)	-483.6629	68.895	-7.020	0.000	-625.024	-342.301
np.log(Girth):np.log(Height)	205.1271	27.591	7.435	0.000	148.515	261.739

Omnibus:	4.947	Durbin-Watson:	1.854
Prob(Omnibus):	0.084	Jarque-Bera (JB):	3.544
Skew:	-0.799	Prob(JB):	0.170
Kurtosis:	3.435	Cond. No.	7.02e+03

The best models based on the Adjusted R-Squared Values appears to the be the log transformed model with variable interactions. This means that as girth increases the height increases the same percentage as does their product.

Part C

1. Without Weight:

without weight
no interactions

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.741
Model:	OLS	Adj. R-squared:	0.723
Method:	Least Squares	F-statistic:	41.42
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	3.16e-09
Time:	19:29:25	Log-Likelihood:	-80.781
No. Observations:	32	AIC:	167.6
Df Residuals:	29	BIC:	172.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	36.9083	2.191	16.847	0.000	32.428	41.389
hp	-0.0191	0.015	-1.275	0.213	-0.050	0.012
cyl	-2.2647	0.576	-3.933	0.000	-3.443	-1.087

Omnibus:	1.178	Durbin-Watson:	1.667
Prob(Omnibus):	0.555	Jarque-Bera (JB):	1.092
Skew:	0.411	Prob(JB):	0.579
Kurtosis:	2.623	Cond. No.	645.

interactions

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.780
Model:	OLS	Adj. R-squared:	0.757
Method:	Least Squares	F-statistic:	33.11
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	2.39e-09
Time:	19:29:25	Log-Likelihood:	-78.143
No. Observations:	32	AIC:	164.3
Df Residuals:	28	BIC:	170.1
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	50.7512	6.512	7.794	0.000	37.413	64.090
hp	-0.1707	0.069	-2.470	0.020	-0.312	-0.029
cyl	-4.1191	0.988	-4.168	0.000	-6.143	-2.095
hp:cyl	0.0197	0.009	2.240	0.033	0.002	0.038

Omnibus:	0.605	Durbin-Watson:	1.767
Prob(Omnibus):	0.739	Jarque-Bera (JB):	0.570
Skew:	0.292	Prob(JB):	0.752
Kurtosis:	2.707	Cond. No.	1.52e+04

With Weight

no interactions

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.843
Model:	OLS	Adj. R-squared:	0.826
Method:	Least Squares	F-statistic:	50.17
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	2.18e-11
Time:	19:29:25	Log-Likelihood:	-72.738
No. Observations:	32	AIC:	153.5
Df Residuals:	28	BIC:	159.3
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	38.7518	1.787	21.687	0.000	35.092	42.412
hp	-0.0180	0.012	-1.519	0.140	-0.042	0.006
cyl	-0.9416	0.551	-1.709	0.098	-2.070	0.187
wt	-3.1670	0.741	-4.276	0.000	-4.684	-1.650

Omnibus:	5.519	Durbin-Watson:	1.644
Prob(Omnibus):	0.063	Jarque-Bera (JB):	4.220
Skew:	0.870	Prob(JB):	0.121
Kurtosis:	3.369	Cond. No.	662.

interactions

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.896
Model:	OLS	Adj. R-squared:	0.866
Method:	Least Squares	F-statistic:	29.52
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	2.64e-10
Time:	19:29:25	Log-Likelihood:	-66.173
No. Observations:	32	AIC:	148.3
Df Residuals:	24	BIC:	160.1
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	43.9654	30.321	1.450	0.160	-18.613	106.544
hp	-0.0259	0.240	-0.108	0.915	-0.521	0.469
cyl	-0.5219	6.337	-0.082	0.935	-13.601	12.558
hp:cyl	-0.0057	0.039	-0.148	0.884	-0.085	0.074
wt	-2.2552	10.684	-0.211	0.835	-24.306	19.796
hp:wt	-0.0367	0.094	-0.392	0.699	-0.230	0.157
cyl:wt	-0.4299	1.991	-0.216	0.831	-4.538	3.678
hp:cyl:wt	0.0065	0.014	0.475	0.639	-0.022	0.035

Omnibus:	1.403	Durbin-Watson:	2.219
Prob(Omnibus):	0.496	Jarque-Bera (JB):	1.216
Skew:	0.307	Prob(JB):	0.544
Kurtosis:	2.269	Cond. No.	4.14e+05

2.

OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.788
Model:	OLS	Adj. R-squared:	0.748
Method:	Least Squares	F-statistic:	19.35
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	5.02e-08
Time:	19:23:49	Log-Likelihood:	-77.542
No. Observations:	32	AIC:	167.1
Df Residuals:	26	BIC:	175.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	35.9830	3.889	9.252	0.000	27.989	43.977
cyl[T.6]	-15.3092	7.435	-2.059	0.050	-30.591	-0.027
cyl[T.8]	-17.9030	5.260	-3.404	0.002	-28.714	-7.092
hp	-0.1128	0.046	-2.465	0.021	-0.207	-0.019
hp:cyl[T.6]	0.1052	0.068	1.536	0.137	-0.036	0.246
hp:cyl[T.8]	0.0985	0.049	2.026	0.053	-0.001	0.198

Omnibus:	1.589	Durbin-Watson:	1.813
Prob(Omnibus):	0.452	Jarque-Bera (JB):	0.862
Skew:	0.393	Prob(JB):	0.650
Kurtosis:	3.165	Cond. No.	3.25e+03

3. Here are the options to be selected from

	hp	cyl	wt
0	100.0	4.0	2.1
1	210.0	8.0	3.9
2	200.0	6.0	2.9

. Predicting using the model you receive

0	25.101601
1	15.114110
2	15.585236

. Based upon the criteria car 1 should be chosen as it should get 25.1 MPG.

Part D

2. Null Fit

OLS Regression Results

Dep. Variable:	chol	R-squared:	0.000
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	nan
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	nan
Time:	19:36:08	Log-Likelihood:	-1988.8
No. Observations:	383	AIC:	3980.
Df Residuals:	382	BIC:	3983.
Df Model:	0		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	208.0601	2.228	93.398	0.000	203.680	212.440

Omnibus:	79.088	Durbin-Watson:	1.979
Prob(Omnibus):	0.000	Jarque-Bera (JB):	201.558
Skew:	0.998	Prob(JB):	1.71e-44
Kurtosis:	5.940	Cond. No.	1.00

Full Fit

OLS Regression Results

Dep. Variable:	chol	R-squared:	0.175
Model:	OLS	Adj. R-squared:	0.105
Method:	Least Squares	F-statistic:	2.492
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	4.27e-05
Time:	19:36:08	Log-Likelihood:	-1951.9
No. Observations:	383	AIC:	3966.
Df Residuals:	352	BIC:	4088.
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1893.5088	1996.951	0.948	0.344	-2033.947	5820.965
gender[T.male]	-1149.1208	865.751	-1.327	0.185	-2851.816	553.575
frame[T.medium]	-525.5859	638.136	-0.824	0.411	-1780.625	729.453
frame[T.small]	-300.3202	682.110	-0.440	0.660	-1641.843	1041.203
location[T.Louisa]	4.2484	4.600	0.924	0.356	-4.798	13.295
gender[T.male]:frame[T.medium]	1226.1281	993.674	1.234	0.218	-728.157	3180.413
gender[T.male]:frame[T.small]	921.8613	1055.892	0.873	0.383	-1154.789	2998.512
age	-11.6646	11.049	-1.056	0.292	-33.394	10.065
age:gender[T.male]	21.3009	15.932	1.337	0.182	-10.034	52.636
age:frame[T.medium]	10.2575	12.495	0.821	0.412	-14.316	34.831
age:frame[T.small]	8.7479	13.705	0.638	0.524	-18.207	35.703
age:gender[T.male]:frame[T.medium]	-17.3682	19.138	-0.908	0.365	-55.008	20.271
age:gender[T.male]:frame[T.small]	-15.8870	19.824	-0.801	0.423	-54.875	23.101
height	-29.8991	30.737	-0.973	0.331	-90.351	30.553
gender[T.male]:height	18.7630	12.954	1.448	0.148	-6.713	44.239
height:frame[T.medium]	8.6497	9.963	0.868	0.386	-10.945	28.244
height:frame[T.small]	4.7488	10.645	0.446	0.656	-16.186	25.684
gender[T.male]:height:frame[T.medium]	-19.3167	14.848	-1.301	0.194	-48.518	9.885
gender[T.male]:height:frame[T.small]	-14.5704	15.715	-0.927	0.354	-45.478	16.337
age:height	0.2026	0.174	1.164	0.245	-0.140	0.545
age:gender[T.male]:height	-0.3541	0.241	-1.468	0.143	-0.828	0.120
age:height:frame[T.medium]	-0.1662	0.197	-0.843	0.400	-0.554	0.222
age:height:frame[T.small]	-0.1373	0.216	-0.636	0.525	-0.562	0.287
age:gender[T.male]:height:frame[T.medium]	0.2812	0.288	0.975	0.330	-0.286	0.848
age:gender[T.male]:height:frame[T.small]	0.2542	0.299	0.849	0.396	-0.334	0.843
waist	-33.3324	50.356	-0.662	0.508	-132.368	65.703
waist:height	0.5967	0.770	0.775	0.439	-0.918	2.111
hip	-32.6065	46.725	-0.698	0.486	-124.502	59.289
waist:hip	0.8937	1.108	0.807	0.420	-1.285	3.072
height:hip	0.5516	0.719	0.768	0.443	-0.862	1.965
waist:height:hip	-0.0153	0.017	-0.898	0.370	-0.049	0.018

Omnibus:	74.944	Durbin-Watson:	1.888
Prob(Omnibus):	0.000	Jarque-Bera (JB):	212.181
Skew:	0.909	Prob(JB):	8.42e-47
Kurtosis:	6.160	Cond. No.	1.11e+08

3. Forward Selection

```
forward selected: chol ~ age*gender*frame
                            OLS Regression Results
==============================================================================
Dep. Variable:                   chol   R-squared:                       0.139
Model:                            OLS   Adj. R-squared:                  0.113
Method:                 Least Squares   F-statistic:                     5.441
Date:                Wed, 18 Apr 2018   Prob (F-statistic):           4.78e-08
Time:                        19:36:09   Log-Likelihood:                -1960.1
No. Observations:                 383   AIC:                             3944.
Df Residuals:                     371   BIC:                             3992.
Df Model:                          11
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      149.2487     20.984      7.112      0.000     107.986     190.511
gender[T.male]                  95.8396     28.636      3.347      0.001      39.531     152.149
frame[T.medium]                 23.0881     24.009      0.962      0.337     -24.122      70.298
frame[T.small]                  -8.5959     25.392     -0.339      0.735     -58.527      41.335
gender[T.male]:frame[T.medium] -57.0444     35.137     -1.623      0.105    -126.138      12.049
gender[T.male]:frame[T.small]  -54.6747     38.105     -1.435      0.152    -129.604      20.255
age                              1.1568      0.383      3.023      0.003       0.404       1.909
age:gender[T.male]              -1.8694      0.519     -3.603      0.000      -2.890      -0.849
age:frame[T.medium]             -0.2269      0.452     -0.502      0.616      -1.116       0.662
age:frame[T.small]               0.2078      0.496      0.419      0.675      -0.768       1.183
age:gender[T.male]:frame[T.medium]  0.9898      0.670      1.477      0.141      -0.328       2.308
age:gender[T.male]:frame[T.small]   0.7749      0.764      1.014      0.311      -0.727       2.277
==============================================================================
Omnibus:                       83.934   Durbin-Watson:                   1.933
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              254.402
Skew:                           0.994   Prob(JB):                     5.72e-56
Kurtosis:                       6.463   Cond. No.                     1.79e+03
==============================================================================
```

4. Backward Selection

backward selected: chol ~ waist*hip+age*gender

OLS Regression Results

Dep. Variable:	chol	R-squared:	0.126
Model:	OLS	Adj. R-squared:	0.112
Method:	Least Squares	F-statistic:	8.996
Date:	Wed, 18 Apr 2018	Prob (F-statistic):	3.40e-09
Time:	19:36:09	Log-Likelihood:	-1963.1
No. Observations:	383	AIC:	3940.
Df Residuals:	376	BIC:	3968.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-82.8060	93.994	-0.881	0.379	-267.627	102.015
gender[T.male]	55.0415	13.371	4.116	0.000	28.750	81.333
waist	6.6486	2.379	2.795	0.005	1.972	11.325
hip	4.7751	2.321	2.058	0.040	0.212	9.338
waist:hip	-0.1280	0.053	-2.425	0.016	-0.232	-0.024
age	1.0236	0.171	5.980	0.000	0.687	1.360
age:gender[T.male]	-1.2626	0.268	-4.706	0.000	-1.790	-0.735

Omnibus:	83.281	Durbin-Watson:	1.934
Prob(Omnibus):	0.000	Jarque-Bera (JB):	245.755
Skew:	0.996	Prob(JB):	4.32e-54
Kurtosis:	6.382	Cond. No.	7.72e+04

I did get a different model. By starting a full model and removing you get a model that has the most number of significant predictors, where as using the forward select model you are trying to find a solution with the least number of predictors.