SAS EXAM 2:

a)

```
data;
input age gender married incomeC healthC childC LifeSatC ses smoke spirit
finish LifeSat income;


datalines;
16    0     0     0     38    0     17    17    1     30    1     22    26
28    1     0     0     38    0     16    21    1     39    1     20    15
;

ods html close;
ods html;

proc rsquare adjrsq mse cp;
model LifeSat=age gender married incomeC healthC childC LifeSatC ses smoke
spirit finish income;
run;
```

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables in Model
6	0.8032	0.7123	7.8040	19.61222	gender childC LifeSatC smoke finish income
7	0.8279	0.7276	8.0659	18.57132	gender incomeC childC LifeSatC smoke finish income
8	0.8531	0.7463	8.2992	17.29322	gender incomeC healthC childC LifeSatC smoke finish income
9	0.8898	0.7906	7.7268	14.27130	gender incomeC healthC childC LifeSatC ses smoke finish income
12	0.9002	0.7291	13.0000	18.46994	age gender married incomeC healthC childC LifeSatC ses smoke spirit finish income

The chosen 4 models have a C(p) less than the full model at 13, while also showing better values in all four categories compared to the other models observed.

b)

```
…
ods html close;
ods html;

proc reg;
model LifeSat=gender childC LifeSatC smoke finish income/cli p;
run;

proc reg;
model LifeSat=gender incomeC childC LifeSatC smoke finish income/cli p;

proc reg;
model LifeSat=gender incomeC healthC childC LifeSatC smoke finish income/cli
p;

proc reg;
model LifeStat=gender incomeC healthC childC LifeSatC ses smoke finish
income/cli p;

proc reg;
model LifeSat=age gender married incomeC healthC childC LifeSatC ses smoke
spirit finish income/cli p;
run;
```

RESULTS

MODEL 1 PRESS:534.94649
MODEL 2 PRESS:565.28672
MODEL 3 PRESS:563.28284
MODEL 4 PRESS:664.43687
MODEL 5 (FULL) PRESS:1449.79044

c)

CALCULATIONS TO OBTAIN R SQUARE PREDICTION:

MODEL 1: 1 – (534.94649)/(1295.2) = .58698

MODEL 2: 1 – (565.28672) /(1295.2) = .56355

MODEL 3: 1 – (563.28284) /(1295.2) = .5610

MODEL 4: 1 – (664.43687) /(1295.2) = .48700

MODEL 5: 1 – (1449.79044) /(1295.2) = 0

d) TABLE TO COMPARE MODELS:

Models	Adj R sq	MSE	Cp	R sq pred	PRESS
Gender, childC, LifeSatC, smoke, finish. Income (6)	.7123	19.6122	7.8040	.58698	534.94649
Gender, incomeC, childC, LifeSatC, smoke, finish. Income (7)	.7276	18.57132	8.0659	.56355	565.28672
Gender, incomeC, health, childC, LifeSatC, smoke, finish. Income (8)	.7463	17.27130	8.2992	.5610	563.28284
Gender, incomeC, health, childC, LifeSatC, ses, smoke, finish. Income (9)	.7906	14.27130	7.7268	.48700	664.43687
Age, gender, married, income, health, childC, LifeSatC, ses, smoke, spirit, finish, income (FULL)	.7291	18.46994	13	0	1449.79044

e) The best model is MODEL 1. In terms of the fit, we see that it has a slightly worse MSE and adj R sq than the rest of the models, but it makes up for it with the second lowest Cp. Where the model really separates itself from the rest is in its prediction, with the highest R sq pred and PRESS. Overall, the slightly worse fit and significantly higher prediction makes MODEL 1 the best combined fit.

f)

```
…
ods html close;
ods html;
proc stepwise;
model LifeSat=age gender married incomeC healthC childC LifeSatC ses smoke
spirit finish income;
run;
proc rsquare adjrsq mse cp;
model LifeSat=age gender married incomeC healthC childC LifeSatC ses smoke
spirit finish income;
run;
proc reg;
model LifeSat=gender healthC childC LifeSatC spirit/ cli p;
run;
```

The following chart shows the stepwise regression best model compared to my chosen best model:

Models	Adj R sq	MSE	Cp	R sq pred	PRESS
Stepwise Model: gender, healthC, childC, LifeSatC, spirit (5)	.6096	26.61554	12.1743	.42422	745.75498
MY MODEL: Gender, childC, LifeSatC, smoke, finish, income (6)	.7123	19.6122	7.8040	.58698	534.94649

Comparing the two models, my chosen model has a higher Adj R sq, a lower MSE, AND a lower Cp. In addition, my chosen model has a higher R sq pred AND a lower PRESS. These numbers show that my chosen model is better in terms of fit AND prediction. I believe this is a reason why we were taught to not always trust stepwise regression, because it does not always select the best model.

g)

```
proc stepwise;
model income=age gender married incomeC healthC childC LifeSatC ses smoke
spirit finish LifeSat;
run;
```

Dependent Variable: income

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	finish		1	0.2359	0.2359	15.8942	5.56	0.0299
2	LifeSat		2	0.2073	0.4432	9.2396	6.33	0.0222
3	gender		3	0.1226	0.5659	6.1216	4.52	0.0494
4	smoke		4	0.0579	0.6237	5.7051	2.31	0.1495
5	healthC		5	0.0682	0.6920	4.8563	3.10	0.1000

Dependent Variable: LifeSat

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	LifeSatC		1	0.2334	0.2334	37.7544	5.48	0.0309
2	childC		2	0.1636	0.3970	28.2824	4.61	0.0465
3	healthC		3	0.1903	0.5874	16.9369	7.38	0.0152
4	spirit		4	0.0748	0.6622	13.6894	3.32	0.0883
5	gender		5	0.0501	0.7123	12.1743	2.44	0.1406

The best model, according to stepwise regression, for fit and prediction for income is the following:

Income = -163.19005 – 21.33326(gender) + 2.25941(healthC) + 41.87748(smoke) + 39.26514(finish) + 1.88927(LifeSat)

The best model, according to stepwise regression, for fit and prediction for LifeSat is the following:

LifeSat = -19.32422 + 3.69279(gender) + 0.70131(healthC) – 5.99778(childC) + 0.50834(LifeSatC) + 0.18906(spirit)

The results are not that meaningful when trying to compare the two models because we can't always trust that the stepwise regression will always choose the best model for fit and prediction. We would need to repeat steps a-e to truly know if these models are the best. As we saw earlier, the model chosen for LifeSat by the stepwise regression was not superior to the model I chose.

h)

```
…
ods html close;
ods html;
proc corr; var age gender married incomeC healthC childC LifeSatC ses smoke
spirit finish LifeSat income;
run;
```

Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	age	gender	married	incomeC	healthC	childC	LifeSatC	ses	smoke	spirit	finish	LifeSat	income
age	1.00000	-0.00515 0.9828	-0.02921 0.9027	0.61175 0.0042	0.10231 0.6678	0.36759 0.1108	0.15198 0.5224	0.22020 0.3509	0.05969 0.8026	-0.05649 0.8130	0.18755 0.4285	0.01615 0.9461	0.13470 0.5713
gender	-0.00515 0.9828	1.00000	-0.19192 0.4176	-0.45410 0.0443	-0.11471 0.6301	-0.07538 0.7521	-0.07352 0.7580	-0.03966 0.8682	0.15352 0.5181	0.17498 0.4606	0.17408 0.4629	0.20232 0.3923	-0.13757 0.5630
married	-0.02921 0.9027	-0.19192 0.4176	1.00000	0.51519 0.0201	0.30071 0.1976	0.32664 0.1598	0.84884 <.0001	0.30404 0.1925	-0.15352 0.5181	0.00357 0.9881	-0.40618 0.0756	0.47209 0.0356	0.04154 0.8619
incomeC	0.61175 0.0042	-0.45410 0.0443	0.51519 0.0201	1.00000	0.54779 0.0124	0.55591 0.0109	0.57490 0.0080	0.53103 0.0160	-0.31612 0.1745	-0.14434 0.5438	-0.14037 0.5550	0.31686 0.1735	0.30061 0.1978
healthC	0.10231 0.6678	-0.11471 0.6301	0.30071 0.1976	0.54779 0.0124	1.00000	0.42606 0.0611	0.27283 0.2445	0.74729 0.0002	-0.89861 <.0001	-0.32414 0.1633	-0.17809 0.4525	0.39346 0.0861	0.06548 0.7839
childC	0.36759 0.1108	-0.07538 0.7521	0.32664 0.1598	0.55591 0.0109	0.42606 0.0611	1.00000	0.58744 0.0065	0.75376 0.0001	-0.21822 0.3553	-0.33754 0.1455	-0.28868 0.2171	-0.04349 0.8555	-0.12567 0.5976
LifeSatC	0.15198 0.5224	-0.07352 0.7580	0.84884 <.0001	0.57490 0.0080	0.27283 0.2445	0.58744 0.0065	1.00000	0.36221 0.1166	-0.10158 0.6700	-0.06183 0.7957	-0.48378 0.0307	0.48316 0.0309	-0.00014 0.9995
ses	0.22020 0.3509	-0.03966 0.8682	0.30404 0.1925	0.53103 0.0160	0.74729 0.0002	0.75376 0.0001	0.36221 0.1166	1.00000	-0.55637 0.0108	-0.16141 0.4966	-0.14603 0.5390	0.18582 0.4328	-0.00233 0.9922
smoke	0.05969 0.8026	0.15352 0.5181	-0.15352 0.5181	-0.31612 0.1745	-0.89861 <.0001	-0.21822 0.3553	-0.10158 0.6700	-0.55637 0.0108	1.00000	0.28688 0.2201	0.12599 0.5966	-0.36336 0.1153	0.04215 0.8599
spirit	-0.05649 0.8130	0.17498 0.4606	0.00357 0.9881	-0.14434 0.5438	-0.32414 0.1633	-0.33754 0.1455	-0.06183 0.7957	-0.16141 0.4966	0.28688 0.2201	1.00000	0.14360 0.5459	0.27849 0.2345	0.33993 0.1425
finish	0.18755 0.4285	0.17408 0.4629	-0.40618 0.0756	-0.14037 0.5550	-0.17809 0.4525	-0.28868 0.2171	-0.48378 0.0307	-0.14603 0.5390	0.12599 0.5966	0.14360 0.5459	1.00000	-0.21523 0.3621	0.48569 0.0299
LifeSat	0.01615 0.9461	0.20232 0.3923	0.47209 0.0356	0.31686 0.1735	0.39346 0.0861	-0.04349 0.8555	0.48316 0.0309	0.18582 0.4328	-0.36336 0.1153	0.27849 0.2345	-0.21523 0.3621	1.00000	0.34014 0.1423
income	0.13470 0.5713	-0.13757 0.5630	0.04154 0.8619	0.30061 0.1978	0.06548 0.7839	-0.12567 0.5976	-0.00014 0.9995	-0.00233 0.9922	0.04215 0.8599	0.33993 0.1425	0.48569 0.0299	0.34014 0.1423	1.00000

For the x's, we see that Finish is moderately correlated to LifeSatC and Income, -48.37% and 48.57% respectively.

We also see in the x's that ChildC and LifeSatC are moderately-highly correlated at 58.74%. Overall, this is the only minor concern within the x's, so we shouldn't have a problem with multicollinearity.

In terms of how the x's are correlated with the dependent variable LifeSat, we can see that only ChildC has close to zero correlation with Life Satisfaction. In this sense, having this in our model is probably not necessary. The other 4 variables all have a moderate correlation with LifeSat which is good for our model.