(1) Develop a regression model to predict using: all variables; forward selection; backward selection; stepwise; the best subset procedure.

See the following result:

"Simple regression all variables",

"Simple regression forward selection",

"Simple regression backward selection",

"Simple regression stepwise selection",

"Simple regression best subset".

(2) What model do you recommend?

The result "Simple regression backward selection".

The variables are "male_fem" on "pct_u18" and "pct_o65". The model is $\text{male\_fem} = 161.08957 - 1.58719 \times \text{pct\_u18} - 2.24835 \times \text{pct\_o65}$.

As shown in the following chart, the "Pr>F" is "<0.0001", and the "Pr>|t|" are both "<0.0001", which means they are significant. As a result, it is a good model.

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			

Root MSE	27.48579	R-Square	0.1441
Dependent Mean	90.74734	Adj R-Sq	0.1419
Coeff Var	30.28826		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type I	Variance Inflation
Intercept	1	161.08957	6.91759	23.29	<.0001	.	0
PCT_U18	1	-1.58719	0.21987	-7.22	<.0001	0.01659	1.11857
PCT_O65	1	-2.24835	0.20764	-10.83	<.0001	0.12967	1.11857

(3) What is the conclusion regarding the overall significance of the regression model? Why?

The F-test considers the linear relationship between the target variable y and the set of predictors taken as a whole. As shown in the following chart, the "F value" is 66.25 and the "Pr>F" is "<0.0001", which means it is significant. As a result, it is a good model.

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			

(4) What variables are included in your model?

"Pct_u18" and "pct_o65" variable.

(5) What is the average error in prediction?

$$\text{average error} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-m-1}} = \sqrt{\frac{594554}{790-2-1}} = \sqrt{755.46851} = 27.486$$

(6) Which of the predictors belong or do not belong in the model?

"Pct_u18" and "pct_o65" variables belong in the model.

"Tot_pop" and "pc_18_65" variables do not belong in the model.

(7) Suppose that we omit total population from the model and rerun the regression. Explain what will happen to the value of "Rsqr"?

The "R-Square" does not change a lot, because maybe there are correlation between "tot_pop" and some other variables. However, it decrease a little, because more variables might give higher multiple correlation, which means less variables might give lower multiple correlation. The "Adj R-Sq" increased a little, because the number of variables reduced. (The following charts are result "with 'tot_pop' and without 'tot_pop'")

Root MSE	27.49882	R-Square	0.1444
Dependent Mean	90.74734	Adj R-Sq	0.1411
Coeff Var	30.30261		

Root MSE	27.48582	R-Square	0.1441
Dependent Mean	90.74734	Adj R-Sq	0.1419
Coeff Var	30.28829		

(8) Discuss the presence of multicollinearity. Does your model contain multi-collinear variables? How do you know?

Yes, there are multi-collinear variables contained in my model. As shown in the following chart, the "condition index" are bigger than 10, and the "proportion of variation" for some variables are bigger than 0.5.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	161.08957	6.91759	23.29	<.0001	.	0
PCT_U18	1	-1.58719	0.21987	-7.22	<.0001	0.89400	1.11857
PCT_O65	1	-2.24835	0.20764	-10.83	<.0001	0.89400	1.11857

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation		
			Intercept	PCT_U18	PCT_O65
1	2.88882	1.00000	0.00237	0.00383	0.01077
2	0.09870	5.41006	0.00809	0.11234	0.60797
3	0.01248	15.21505	0.98954	0.88384	0.38126