1)

a)

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_i$

b)

```
> SavingsMod=lm(y~x1+x2+x3+x4+x5,data = Savings)
> SavingsMod

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = Savings)

Coefficients:
(Intercept)          x1          x2          x3          x4          x5
 26.3738501  -0.4253347  -1.4870502  -0.0003265  -0.5579129   1.0921893
```

Y=26.3738501 - .4253347x1 - 1.4870502x2 - .0003265x3 - .5579129x4 + 1.0921893x5

c)

```
Y=26.3738501 - .4253347*25 - 1.4870502*5 - .0003265*2000 - .5579129*4 + 1.0921893*5
[1] 10.88153
```

d)

```
> summary(SavingsMod)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = Savings)

Residuals:
   Min     1Q  Median     3Q    Max
-8.2632 -2.2672  0.0002  2.1849  8.7561

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.3738501  7.9297806   3.326  0.00179 **
x1          -0.4253347  0.1527564  -2.784  0.00788 **
x2          -1.4870502  1.1212637  -1.326  0.19161
x3          -0.0003265  0.0009356  -0.349  0.72874
x4          -0.5579129  1.2854453  -0.434  0.66639
x5           1.0921893  1.4337971   0.762  0.45028
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.821 on 44 degrees of freedom
Multiple R-squared: 0.3471, Adjusted R-squared: 0.2729
F-statistic: 4.678 on 5 and 44 DF, p-value: 0.001643

The overall model is statisticallyt significant. Since it is statistically significant, there is evidence that at least one independent variable has an effect on the response variable.

e)
Based off the R-squared value, the linear relationship does not appear to be strong.

f)
> summary(SavingsMod)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = Savings)

Residuals:
   Min   1Q Median   3Q   Max
-8.2632 -2.2672 0.0002 2.1849 8.7561

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept) 26.3738501 7.9297806   3.326 0.00179 **
x1   -0.4253347 0.1527564 -2.784 0.00788 **
x2   -1.4870502 1.1212637 -1.326 0.19161
x3   -0.0003265 0.0009356 -0.349 0.72874
x4   -0.5579129 1.2854453 -0.434 0.66639
x5   1.0921893 1.4337971   0.762 0.45028

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.821 on 44 degrees of freedom
Multiple R-squared: 0.3471, Adjusted R-squared: 0.2729
F-statistic: 4.678 on 5 and 44 DF, p-value: 0.001643

Based of the estimates, it seems like the first hypothesis is correct since the the first two variables have negative coefficients. For the second hypothesis, the variable has a very high p-value

making it not likely to have an effect and it is not statistically significant. The third hypothesis suggest that the fourth and fifth variable would be similar to their relationship to the dependant variable. Looking at their correlation, they do have a similar correlation to the dependanet variable. x5's coefficient being close to 1 and its p-value suggest the variable suggest x5 is not contributing much to this model.

g)

H0:β2=0

H1:β2≠0

```
> SavingsModX1UX2=lm(y~x1+x2,data = Savings)
> SavingsAnovaX1X2=anova(SavingsModX1UX2)
> SavingsAnovaX1X2
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 204.12 204.118 13.2112 0.0006879 ***
x2         1  53.34  53.343  3.4525 0.0694254 .
Residuals 47 726.17  15.450
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> qf(.95,1,47)
[1] 4.0471
```

Since F* is higher than F for x1 we reject the null, but fail to reject for x2 meaning it can be removed. So the variables x1 and x2 do not have an equal contribution.

h)

```
> cor(Savings)
           y          x1          x2          x3          x4          x5
y   1.0000000 -0.45553809  0.31652112  0.2203435  0.30478716  0.33622223
x1 -0.4555381  1.00000000 -0.90847871 -0.7561760 -0.04782569 -0.07565359
x2  0.3165211 -0.90847871  1.00000000  0.7869882  0.02532138  0.03418780
x3  0.2203435 -0.75617603  0.78698824  1.0000000 -0.12948721 -0.11675536
x4  0.3047872 -0.04782569  0.02532138 -0.1294872  1.00000000  0.98786727
x5  0.3362222 -0.07565359  0.03418780 -0.1167554  0.98786727  1.00000000
```

X4 and x5 seem to have the highest correlation which makes sense considering the problem stated they are derived from a similar formula.

i)

```
SavingsModX1234=lm(y~x1+x2+x3+x4,data = Savings)
> summary(SavingsModX1234)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = Savings)

Residuals:
   Min      1Q  Median      3Q     Max
-8.2422 -2.6858 -0.2487  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5662295  7.3545025   3.884 0.000334 ***
x1          -0.4611959  0.1446418  -3.189 0.002603 **
x2          -1.6914510  1.0835935  -1.561 0.125538
x3          -0.0003370  0.0009311  -0.362 0.719078
x4           0.4096889  0.1961967   2.088 0.042474 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,  Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007903
```

```
data Saving;
infile 'saving.txt' firstobs=2;
input y x1 x2 x3 x4 x5;
run;
proc print data=Saving;
run;
proc reg data=Saving;
model y= x1 x2 x3 x4 x5/selection=Rsquare AdjRsq Cp AIC SBC best=3
Run;
```

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	SBC	Variables in Model
1	0.2075	0.1910	7.4043	141.3322	145.15622	x1
1	0.1130	0.0946	13.7704	146.9632	150.78722	x5
1	0.1002	0.0814	14.6370	147.6829	151.50696	x2
2	0.2991	0.2693	3.2327	137.1919	142.92798	x1 x5
2	0.2878	0.2575	3.9948	137.9923	143.72834	x1 x4
2	0.2617	0.2303	5.7498	139.7879	145.52399	x1 x2
3	0.3426	0.2997	2.3029	135.9901	143.63824	x1 x2 x5
3	0.3365	0.2933	2.7100	136.4475	144.09563	x1 x2 x4
3	0.3119	0.2670	4.3708	138.2711	145.91916	x1 x3 x5
4	0.3453	0.2871	4.1218	137.7853	147.34546	x1 x2 x4 x5
4	0.3443	0.2860	4.1884	137.8607	147.42083	x1 x2 x3 x5
4	0.3385	0.2797	4.5803	138.3022	147.86230	x1 x2 x3 x4
5	0.3471	0.2729	6.0000	139.6471	151.11925	x1 x2 x3 x4 x5

Like mentioned previously, x5 does not seem to be contributing a lot to this model. When looking at a model without x5, it actually seems that x4 becoming statistically significant. Furthermore, when comparing all the models, it seems that both x4 and x5 should not be together unless x3 is the only absent variable

j)

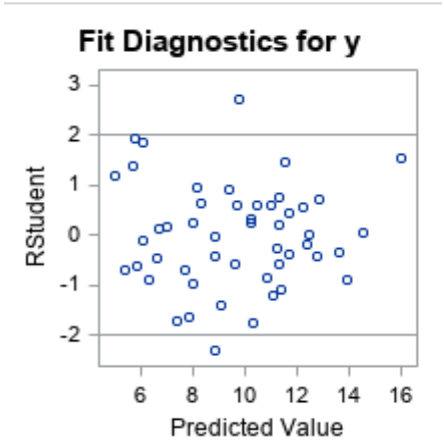
```
> SavingsAnova=anova(SavingsMod)
> SavingsAnova
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 204.12 204.118 13.9841 0.0005294 ***
x2         1  53.34  53.343  3.6545 0.0624439 .
x3         1  12.40  12.404  0.8498 0.3616368
x4         1  63.05  63.052  4.3197 0.0435375 *
x5         1   8.47   8.470  0.5803 0.4502755
Residuals 44 642.24  14.596
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> SSE=SavingsAnova$`Sum Sq`[6]
> SSRX5gX1234=SavingsAnovaX1234$`Sum Sq`[5]-SSE
> SSRX5gX1234
[1] 8.46967
> R2=SSRX5gX1234/SavingsAnovaX1234$`Sum Sq`[5]
> R2
[1] 0.01301601
```

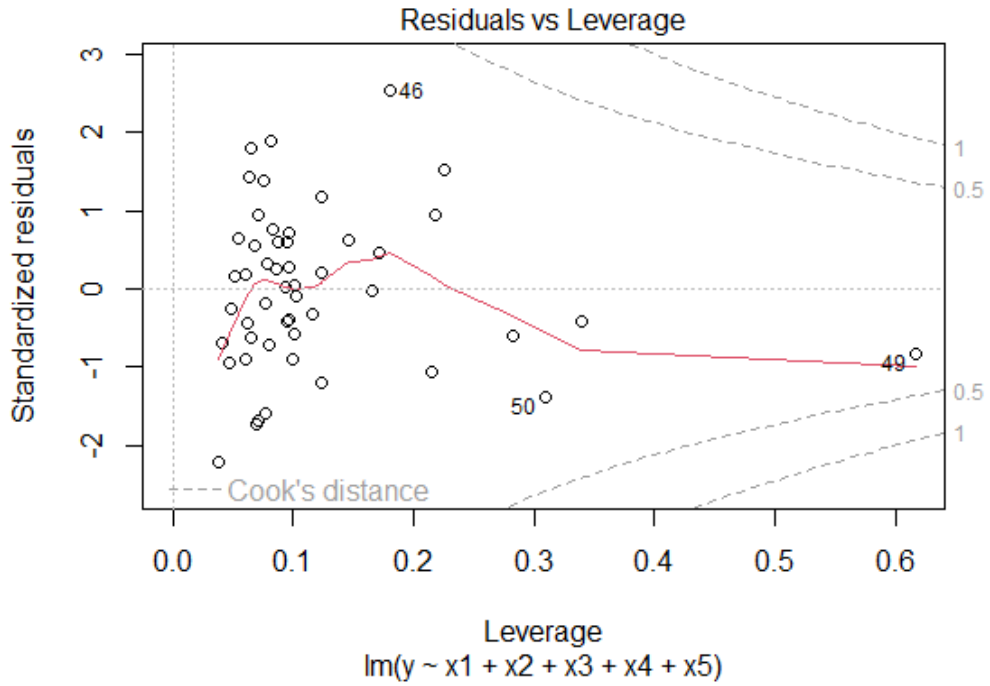
Based off the R2, X5 does not provide much more information than the first four variables already do.

k)



Fairly symmetrical on both sides of the 0, with few outliers.

l)



No points fall in the lines so the model contains no real influential points. The close point to falling under this is data point 46, 49, and 50.

m)

32	2.02	7.8847	1.0631	-5.8647	3.670	-1.598	0.036
33	12.70	5.7930	1.0874	6.9070	3.663	1.886	0.052
34	12.78	6.1387	0.9695	6.6413	3.695	1.797	0.037
35	12.49	13.6580	1.3012	-1.1680	3.592	-0.325	0.002
36	11.14	10.2355	1.1209	0.9045	3.652	0.248	0.001
37	13.30	11.7091	1.5817	1.5909	3.478	0.457	0.007
38	11.77	12.4240	1.0626	-0.6540	3.670	-0.178	0.000
39	6.86	11.1216	1.3453	-4.2616	3.576	-1.192	0.034
40	14.13	11.3540	1.1038	2.7760	3.658	0.759	0.009
41	5.13	7.7213	0.7669	-2.5913	3.743	-0.692	0.003
42	2.81	5.4038	1.0840	-2.5938	3.664	-0.708	0.007
43	7.81	11.4178	1.7733	-3.6078	3.384	-1.066	0.052
44	7.56	8.8802	2.2239	-1.3202	3.107	-0.425	0.015
45	9.22	5.0226	1.3450	4.1974	3.576	1.174	0.032
46	18.56	9.8039	1.6262	8.7561	3.457	2.533	0.237
47	7.72	9.6433	2.0303	-1.9233	3.236	-0.594	0.023
48	9.24	11.3261	1.2183	-2.0861	3.621	-0.576	0.006
49	8.89	10.8697	3.0003	-1.9797	2.365	-0.837	0.188
50	4.71	9.1198	2.1266	-4.4098	3.174	-1.389	0.144

Taking a look at the output statistics, it further supports data points 46,49,50 being the most influential points. They all have higher Cook's D. Using a .05 alpha level, the critical value comes out to around 2.704, which only data point 46 comes close to that from the studentized residuals.

2)

```
data Bweights;
infile 'Bweights.txt' dlm='09'x obs=1000;
input bgender bhead   blengthbwt      delwt gaweek fincome frace mrace malform enarche
mheight momage        parity ppbmi ppwt smoken evrsmok;
run;
proc print data=Bweights;
run;
proc reg data=Bweights;
model bwt=bgender bhead     blengthdelwt gaweek fincome frace mrace malform enarche
mheight momage        parity ppbmi ppwt smoken evrsmok/selection=Rsquare AdjRsq Cp AIC
SBC best=1;
Run;
proc reg data= Bweights;
model bwt= bgender bhead blength delwt gaweek mrace momage ppbmi evrsmok;
run;
```

Results Viewer - sashtml

in Model	R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
1	0.5766	0.5762	538.5383	-649.7078	-639.89226	bhead
2	0.7006	0.7000	91.3445	-994.0399	-979.31667	bhead blength
3	0.7111	0.7102	55.3168	-1027.7059	-1008.07489	bhead blength mrace
4	0.7177	0.7165	33.2876	-1048.9168	-1024.37800	bhead blength gaweek mrace
5	0.7208	0.7194	23.9045	-1058.1032	-1028.65663	bhead blength delwt gaweek mrace
6	0.7230	0.7214	17.9086	-1064.0364	-1029.68213	bgender bhead blength delwt gaweek mrace
7	0.7250	0.7230	12.8533	-1069.0890	-1029.82698	bgender bhead blength delwt gaweek mrace ppbmi
8	0.7264	0.7242	9.7709	-1072.2005	-1028.03067	bgender bhead blength delwt gaweek mrace ppbmi evrsmok
9	0.7277	0.7252	7.1498	-1074.8708	-1025.79328	bgender bhead blength delwt gaweek mrace momage ppbmi evrsmok
10	0.7281	0.7253	7.5439	-1074.4989	-1020.51364	bgender bhead blength delwt gaweek mrace momage ppbmi smoken evrsmok
11	0.7284	0.7254	8.3625	-1073.6984	-1014.80529	bgender bhead blength delwt gaweek mrace enarche momage ppbmi smoken evrsmok
12	0.7287	0.7254	9.3318	-1072.7461	-1008.94524	bgender bhead blength delwt gaweek fincome mrace enarche momage ppbmi smoken evrsmok
13	0.7289	0.7253	10.7073	-1071.3813	-1002.67269	bgender bhead blength delwt gaweek fincome mrace enarche momage parity ppbmi smoken evrsmok
14	0.7290	0.7252	12.1217	-1069.9774	-996.36106	bgender bhead blength delwt gaweek fincome frace mrace enarche momage parity ppbmi smoken evrsmok
15	0.7291	0.7249	14.0025	-1068.0988	-989.57471	bgender bhead blength delwt gaweek fincome frace mrace enarche mheight momage parity ppwt smoken evrsmok
16	0.7291	0.7247	16.0001	-1066.1012	-982.66939	bgender bhead blength delwt gaweek fincome frace mrace enarche mheight momage parity ppbmi ppwt smoken evrsmok

The model with 9 explanatory variables seems best although the model with 10 would fit as well. The mode with 9 explanatory variables takes the baby's gender, head circumference,length, mother's weight during delivery, gestational age, mother's race, mother's age at delievery, mother's pre prenancy bmi, and if the mother has smoked or not. It has the lowest C(p) and AIC and has just about the highest R-square. Looking at the p-values of the model, one can see that this model is statisticall significant.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: bwt

Number of Observations Read	1000
Number of Observations Used	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	893.93148	99.32572	293.90	<.0001
Error	990	334.58282	0.33796		
Corrected Total	999	1228.51430			

Root MSE	0.58135	R-Square	0.7277
Dependent Mean	6.90937	Adj R-Sq	0.7252
Coeff Var	8.41387		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-12.84733	0.42732	-30.07	<.0001
bgender	1	-0.10879	0.03749	-2.90	0.0038
bhead	1	0.30585	0.01522	20.10	<.0001
blength	1	0.16239	0.00870	18.66	<.0001
delwt	1	0.00502	0.00122	4.11	<.0001
gaweek	1	0.02634	0.00678	3.88	0.0001
mrace	1	-0.09535	0.02136	-4.46	<.0001
momage	1	0.01063	0.00494	2.15	0.0316
ppbmi	1	-0.02237	0.00818	-2.74	0.0063
evrsmok	1	0.08867	0.03868	2.29	0.0221