

Part I. (85 points) Do all calculations in SAS. Use a word processor of your choice to write a report. Insert computer text output and graphics to support what you are saying, but you need to write something that looks like an academic paper — not a pile of computer output. Turn in a hard copy of your HW in class (i.e., don't email me your HW).

(85^{pts})

1. CCHD birth weight: The California Child Health and Development Study involved women on the Kaiser Health plan who received prenatal care and later gave birth in the Kaiser clinics. Approximately 19,000 live-born children were delivered in the 20,500 pregnancies. We consider the subset of the 680 live-born white male infants in the study. Data were collected on a variety of features of the child, the mother, and the father.

The columns in the data set are, from left to right:

- 1) ID
- 2) child's head circumference (inches)
- 3) child's length (inches), y response
- 4) child's birth weight (pounds)
- 5) gestation (weeks)
- 6) maternal age (years)
- 7) maternal smoking (cigarettes/day)
- 8) maternal height (inches)
- 9) maternal pre-pregnancy weight (pounds)
- 10) paternal age (years)
- 11) paternal education (years)
- 12) paternal smoking (cigarettes/day)
- 13) paternal height (inches)

A goal here is to build a multiple regression model to predict child's birth weight (4) from the data on the mother and father (6–13). A reasonable strategy would be to:

1. Examine the relationship between birth weight and the potential predictors.
2. Decide whether any of the variables should be transformed.
3. Perform a backward elimination using the desired response and predictors.
4. Given the selected model, examine the residuals and check for influential cases.
5. Repeat the process, if necessary.

Given your statistical analysis, provide a writeup that includes the following parts.

- (a) (40 pts) A discussion of the process you used to build the regression model, including relevant output (e.g., make sure to do a residual analysis to check model assumptions).

Solution:

Summary of Process: First I will look at the data, considering whether there are unusual observations for any variables, and considering transformations for linear relationships for the x s with y . Next I will perform backward selection, checking assumptions with the final model, then redo backward selection after any modifications (removing influential observations, additional transformations). Once all model assumptions have been met with the reduced model, I will interpret the model and discuss any model limitations.

Looking at the data: The minimum and maximum values in the "Simple Statistics" table below seem reasonable for each variable.

The scatterplots reveals one outlier for maternal pre-pregnancy weight ($x8$).

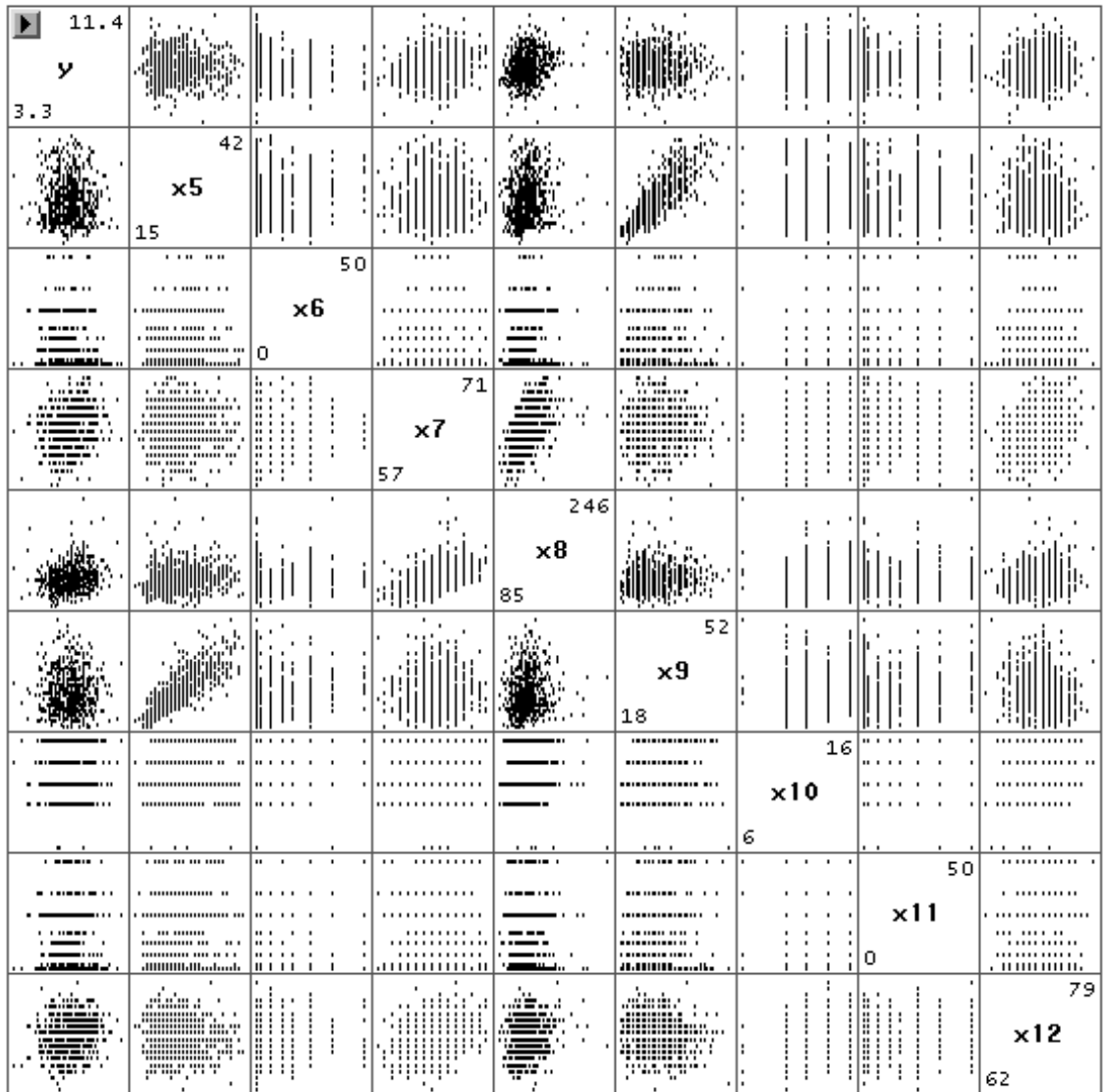
The scatterplots and correlation matrix suggest $x6$, $x7$, $x8$, and $x12$ as weak linear predictors of $y = x3$. There is no strong suggestion for transformations of x s to be linear with $y = x3$ Note that relationships may change with multiple predictors together.

85 pts

40 pts

The CORR Procedure

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
y=x3	680	7.51647	1.09235	5111	3.30000	11.40000	child birth weight (pounds)
x5	680	25.85735	5.46338	17583	15.00000	42.00000	maternal age (years)
x6	680	7.43088	11.27202	5053	0	50.00000	maternal smoking (cigarettes/day)
x7	680	64.43382	2.48323	43815	57.00000	71.00000	maternal height (inches)
x8	680	126.89559	17.87766	86289	85.00000	246.00000	maternal pre-pregnancy weight (pounds)
x9	680	28.80000	6.13313	19584	18.00000	52.00000	paternal age (years)
x10	680	13.37941	2.20259	9098	6.00000	16.00000	paternal education (years)
x11	680	14.43824	14.17030	9818	0	50.00000	paternal smoking (cigarettes/day)
x12	680	70.61912	2.63832	48021	62.00000	79.00000	paternal height (inches)



Pearson Correlation Coefficients, N = 680
Prob > |r| under H0: Rho=0

	y=x3	x5	x6	x7	x8	x9	x10	x11	x12
y=x3 child birth weight (pounds)	1.00000	0.00131 0.9729	-0.17941 <.0001	0.20254 <.0001	0.22158 <.0001	0.01645 0.6685	0.03302 0.3899	-0.02337 0.5430	0.15416 <.0001
x5 maternal age (years)	0.00131 0.9729	1.00000	0.04500 0.2412	0.01749 0.6490	0.11573 0.0025	0.81711 <.0001	0.24059 <.0001	0.01662 0.6653	-0.07111 0.0639
x6 maternal smoking (cigarettes/day)	-0.17941 <.0001	0.04500 0.2412	1.00000	0.02593 0.4996	-0.02576 0.5024	0.02771 0.4707	0.02372 0.5370	0.26171 <.0001	0.01078 0.7791
x7 maternal height (inches)	0.20254 <.0001	0.01749 0.6490	0.02593 0.4996	1.00000	0.49419 0.0048	0.01799 0.6396	0.10799 0.0048	-0.01470 0.7019	0.30333 <.0001
x8 maternal pre-pregnancy weight (pounds)	0.22158 <.0001	0.11573 0.0025	-0.02576 0.5024	0.49419 0.0048	1.00000	0.12399 0.0012	0.00127 0.9736	-0.02747 0.4745	0.16642 <.0001
x9 paternal age (years)	0.01645 0.6685	0.81711 <.0001	0.02771 0.4707	0.01799 0.6396	0.12399 0.0012	1.00000	0.22040 0.3015	0.03968 0.3015	-0.13441 0.0004
x10 paternal education (years)	0.03302 0.3899	0.24059 <.0001	0.02372 0.5370	0.10799 0.0048	0.00127 0.9736	0.22040 <.0001	1.00000	-0.18228 <.0001	0.10778 0.0049
x11 paternal smoking (cigarettes/day)	-0.02337 <.0001	0.01662 0.6653	0.26171 <.0001	-0.01470 0.7019	-0.02747 0.4745	0.03968 0.3015	-0.18228 <.0001	1.00000	0.01365 0.7224
x12 paternal height (inches)	0.15416 <.0001	-0.07111 0.0639	0.01078 0.7791	0.30333 <.0001	0.16642 <.0001	-0.13441 0.0004	0.10778 0.0049	0.01365 0.7224	1.00000

Backward selection: Backward selection results in a model with x6, x7, x8, and x12 all significant at a 0.05 level.

...

Backward Elimination: Step 4

Variable x11 Removed: R-Square = 0.1016 and C(p) = 2.5267

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	82.27891	20.56973	19.07	<.0001
Error	675	727.91662	1.07839		
Corrected Total	679	810.19553			

Variable	Parameter Estimate	Standard Error	Type III SS	F Value	Pr > F
Intercept	0.64388	1.33158	0.25214	0.23	0.6289
x6	-0.01738	0.00354	25.97911	24.09	<.0001
x7	0.04532	0.01912	6.05658	5.62	0.0181
x8	0.00913	0.00257	13.63670	12.65	0.0004
x12	0.04139	0.01586	7.34974	6.82	0.0092

Bounds on condition number: 1.4199, 19.403

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x10	paternal education (years)	7	0.0004	0.1032	7.3301	0.33	0.5658
2	x5	maternal age (years)	6	0.0006	0.1026	5.7647	0.44	0.5097
3	x9	paternal age (years)	5	0.0002	0.1024	3.9045	0.14	0.7083
4	x11	paternal smoking (cigarettes/day)	4	0.0008	0.1016	2.5267	0.62	0.4298

Diagnostics: Observation 506 has larger influence than other observations on the model fit, seen below by the Cook's D. Other diagnostics do not show any issues.

I will remove observation 506 from the data and redo backward selection.

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2	-1	0	1	2	Cook's D
500	6.1000	7.6224	0.0865	-1.5224	1.035	-1.471		**				0.003
501	8.1000	7.4698	0.0640	0.6302	1.036	0.608		*				0.000
502	7.6000	6.5914	0.1226	1.0086	1.031	0.978		*				0.003
503	7.8000	7.1942	0.1089	0.6058	1.033	0.587		*				0.001
504	7.1000	7.2553	0.0783	-0.1553	1.036	-0.150						0.000
505	3.9000	6.9081	0.1208	-3.0081	1.031	-2.917		*****				0.023
506	4.5000	8.0067	0.1926	-3.5067	1.020	-3.436		*****				0.084 <----
507	6.6000	7.3842	0.0541	-0.7842	1.037	-0.756		*				0.000
508	6.0000	7.1626	0.1052	-1.1626	1.033	-1.125		**				0.003
509	9.0000	7.3719	0.0780	1.6281	1.036	1.572		***				0.003
510	8.0000	7.0482	0.0792	0.9518	1.035	0.919		*				0.001

Final model building is completed in the next part.

Program editor contents:

```
* options for sas session;
options ls=79 nodate nocenter;

* part (a) *****;
* read data, create variables, assign labels;
data cchd;
  infile 'F:\Dropbox\UNM\teach\ADA2_stat528\assess\ADA2_HW_03_cchd-birthwt.dat';
  input id x1-x12;
  y = x3;
  label id = 'ID'
        x1 = 'child head circumference (inches)'
        x2 = 'child length (inches)'
        x3 = 'child birth weight (pounds)'
```

```

x4 = 'gestation (weeks)'
x5 = 'maternal age (years)'
x6 = 'maternal smoking (cigarettes/day)'
x7 = 'maternal height (inches)'
x8 = 'maternal pre-pregnancy weight (pounds)'
x9 = 'paternal age (years)'
x10 = 'paternal education (years)'
x11 = 'paternal smoking (cigarettes/day)'
x12 = 'paternal height (inches)'
y = 'child birth weight (pounds)';
run;

* scatterplot matrix;
* USE menu: Solutions > Analysis > Interactive data analysis;
* since proc corr only does x1-x5;
ods html style=journal; * turn on html output;
ods graphics on; * turn on ods graphics;
proc corr data=cchd plots;
var y x5-x12;
run;
ods graphics off;
ods html close;

* regression with backward selection;
proc reg data=cchd;
model y = x5-x12 /selection = backward; * backward selection;
run;

* reduced model with diagnostics;
proc reg data=cchd;
model y = x6 x7 x8 x12 /p r partial;
plot student.*predicted. student.*nqq. cookd.*obs.; * diagnostic plots;
run;

```

(b) (20 pts) A discussion of which maternal and paternal features are useful for predicting child’s birth weight, and whether the variables selected by the backwards elimination procedure “makes sense.” Also, interpret the sign of the regression coefficients in the final model, that is, which predictor variables are positively associated with birth weights (holding the other predictors constant), and which are negatively related.

Solution:

Backward selection without observation 506 (not shown) results in the same model with *x6*, *x7*, *x8*, and *x12* all significant at a 0.05 level conditional on all other variables in the model.

The model fit without observation 506 (below) is highly significant (F-value=20.24, p-value < 0.0001) but only explaining about 10.7% of the variability in child birth weight ($R^2 = 0.1072$). Variables *x6*, *x7*, *x8*, and *x12* are all significant conditional on all other variables in the model.

The REG Procedure

Dependent Variable: y child birth weight (pounds)

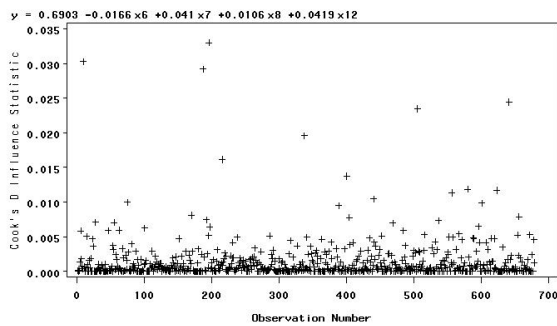
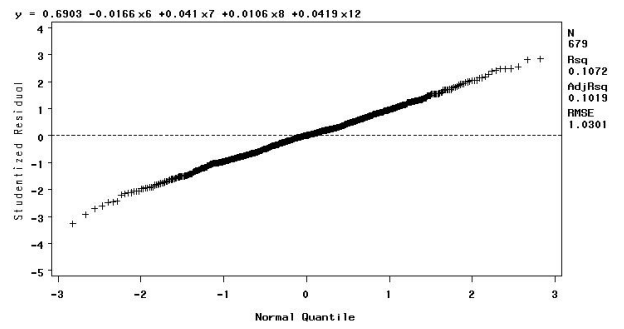
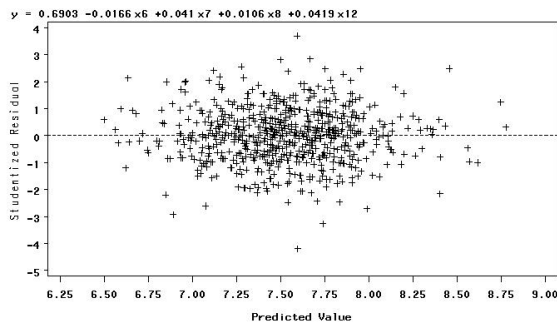
Number of Observations Read 679
 Number of Observations Used 679

Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	<----
Model	4	85.90153	21.47538	20.24	<.0001	
Error	674	715.18151	1.06110			
Corrected Total	678	801.08303				

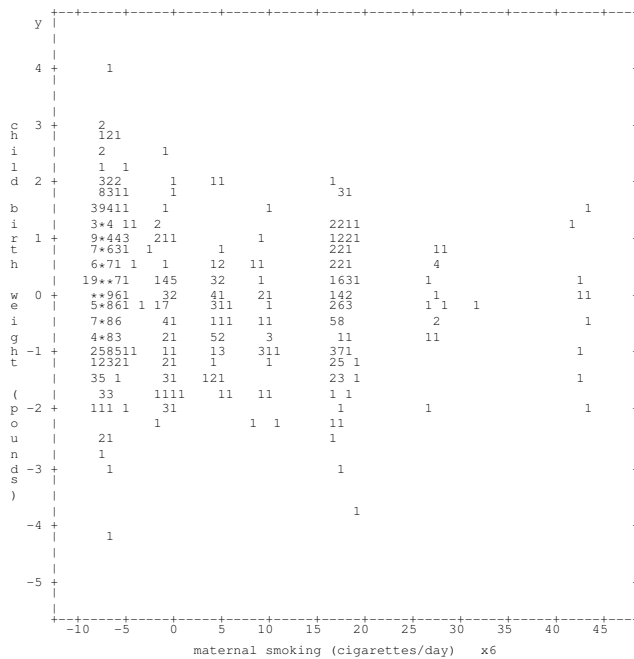
Root MSE 1.03010 R-Square 0.1072 <----
 Dependent Mean 7.52091 Adj R-Sq 0.1019
 Coeff Var 13.69644

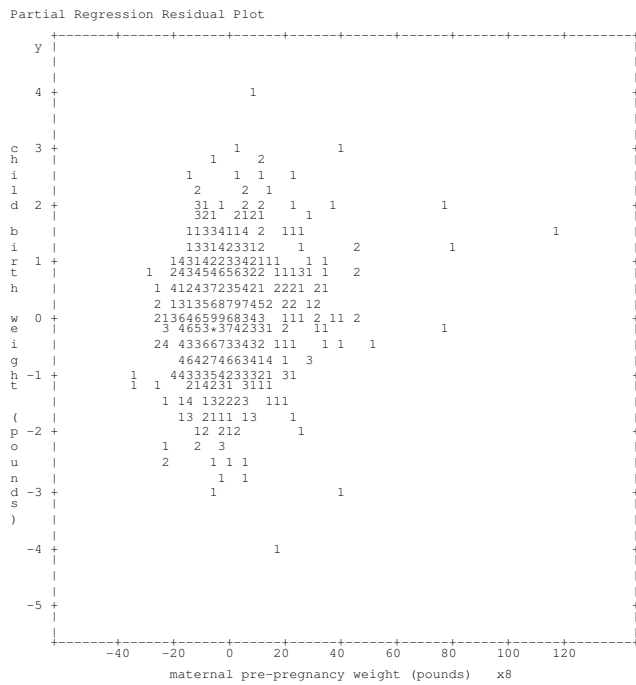
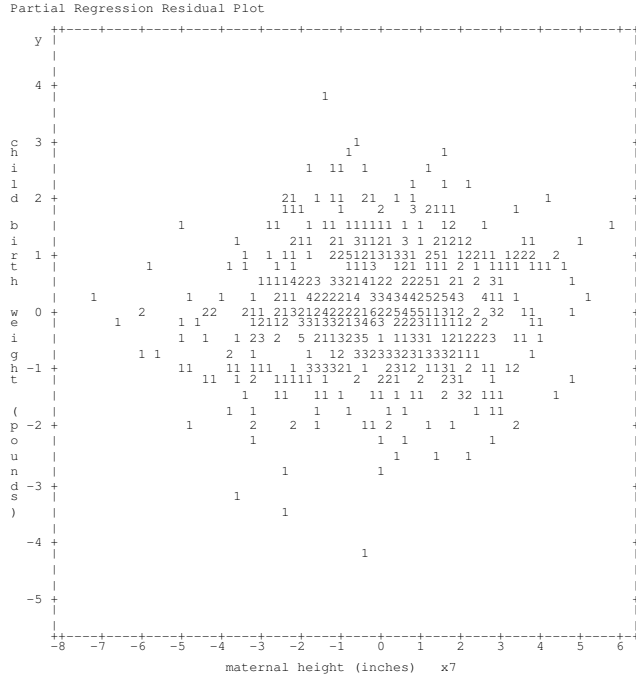
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	Intercept	1	0.69034	1.32093	0.52	0.6014
x6	maternal smoking (cigarettes/day)	1	-0.01655	0.00352	-4.70	<.0001
x7	maternal height (inches)	1	0.04100	0.01901	2.16	0.0314
x8	maternal pre-pregnancy weight (lb)	1	0.01065	0.00258	4.12	<.0001
x12	paternal height (inches)	1	0.04194	0.01573	2.67	0.0079

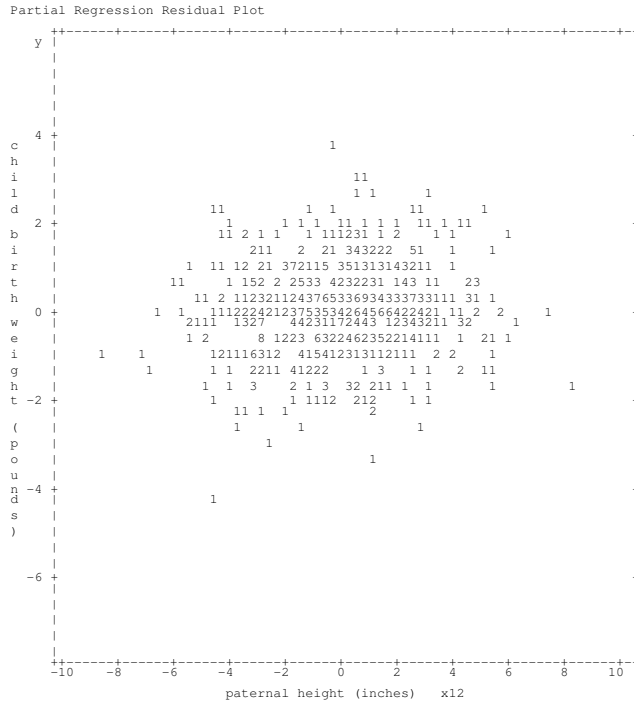
The plots indicate that the residuals do not show any violation of constant variance or normality, and Cook’s distances do not suggest any overly influential observations.



The partial regression residual plots show weak linear relationships for each variable conditional on the others. Variable x_8 shows a potential high-leverage point (high pre-pregnancy weight) that might influence this predictor, but given the Cook's distance plot above, I won't worry too much about this point at this time (though it could be worth redoing the analysis without this point to see whether the selected model changes).







Model interpretation: The final model is

$$y = 0.69 - 0.017x_6 + 0.041x_7 + 0.011x_8 + 0.042x_{12}$$

$$\text{BirthWt(lbs)} = 0.69 - 0.017\text{MatSmoke(cig/day)} + 0.041\text{MatHt(in)} + 0.011\text{MatWt(lbs)} + 0.042\text{PatHt(in)}.$$

The model predicts that, with all other predictors held constant, that a child's birth weight (pounds) will

- decrease 0.017 pounds for each additional cigarette the mother smokes each day,
- increase 0.041 pounds for each inch taller the mother is,
- increase 0.011 pounds for each pound heavier the mother is pre-pregnancy, and
- increase 0.042 pounds for each inch taller the father is.

So maternal cigarette use decreases birth weight and parental physical size increases birth weight. This makes sense to me.

Program editor contents:

```

* remove influential obs 506, and redo backward selection;
data cchd2;
  set cchd;
  if _N_ = 506 then delete;
run;

* regression with backward selection;
proc reg data=cchd2;
  model y = x5-x12 /selection = backward; * backward selection;
run;

* reduced model is the same with obs 506 removed;
* reduced model with diagnostics;
proc reg data=cchd2;
  model y = x6 x7 x8 x12 /p r partial;
  plot student.*predicted. student.*nqq. cookd.*obs.; * diagnostic plots;
run;
    
```

(c) (10 pts) A summary table that includes the important predictors of birth weight, and other features (for example regression coefficients and standard errors, or p-values) that would be useful to report in a scientific paper.

Solution:

The ANOVA and parameter estimate tables above, along with R^2 value are valuable for a report.

- (d) (15 pts) A summary of the analysis, including any potential limitations you might see with your conclusions.

Solution:

We considered a subset of parental variables on their child's birth weight on the Kaiser Health plan who received prenatal care and later gave birth to white males infants in the Kaiser clinics. After removing an influential observation, backward selection chose maternal smoking, height, and pre-pregnancy weight, as well as paternal height as predictors in a model explaining 10.7% of the variability in birth weight. An increase in smoking reduces birth weight while an increase in other factors increases birth weight.