C.4 Multiple Linear Regression

In many developing countries, especially underprivileged sections, last menstrual period is not known and thus gestational age in pregnancy is also not known. The aim of this study is to explore the possibility of predicting gestational age (GA) from the anthropometrical measurements of the newborn. These measurements are birthweight, crown heal length (CHL), head circumference (HC), mid arm circumference (MAC), foot length (FL), hand length (HL), and calf circumference (CaC). The data were obtained on 800 newborns.

Multiple linear regression method using stepwise selection is applied. To build a good model, assumptions required to be tested are linearity, outliers, multicollinearity. Gaussianity of error term will be tested after fitting the model.

The linearity can be explored by drawing the scatter plot.

Command for the scatter plot matrix:

```
GRAPH
/SCATTERPLOT(MATRIX)=GA BirthWeight CHL HC MAC FL HL CaC
/MISSING=LISTWISE.
```

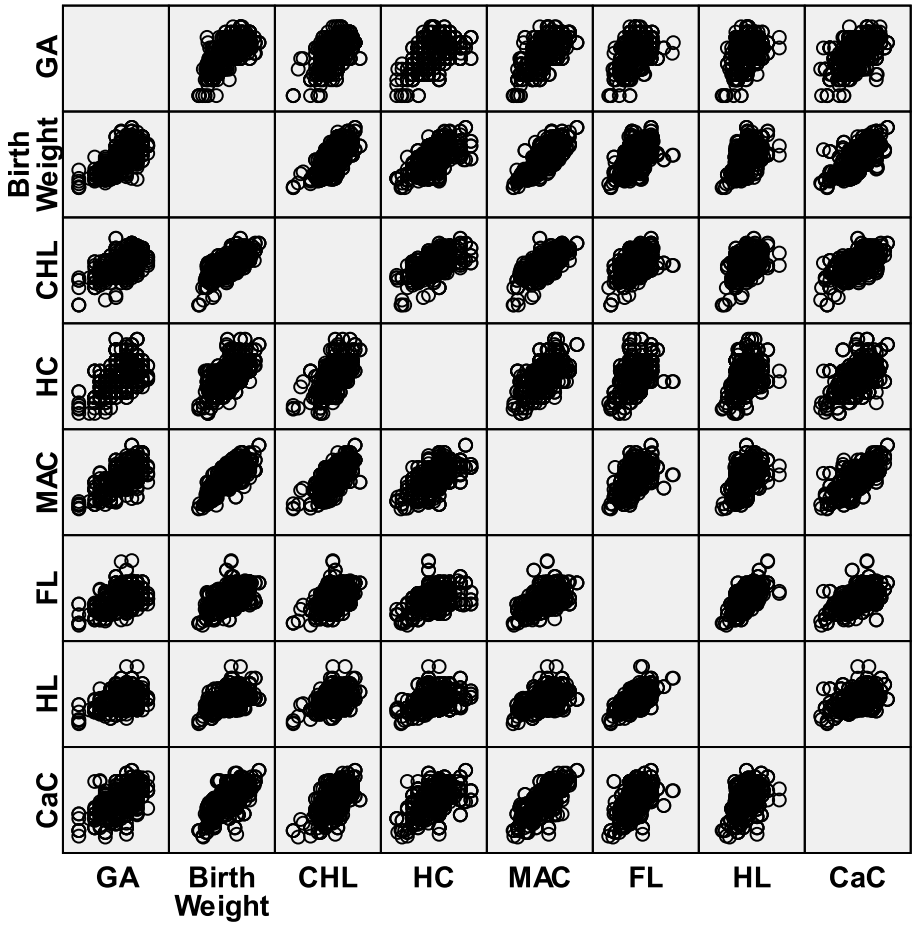

FIGURE C.4(a) Scatter plot in matrix form is shown below.

The scatter matrix gives the visual assessment of linearity. Plots in the first row indicate that most variables are linearly related with the gestational age. BirthWeight may have quadratic relationship as indicated by slightly curved plot but keep it linear for the present example. Apparently there are no outliers. This plot also suggests linear relationship (multicollinearity) among some independents. This is examined in more details slightly later in this output. To further check linearity, compute Pearson correlations.

Command to compute the Pearson correlation:

```
CORRELATIONS
/VARIABLES=GA with BirthWeight CHL MAC FL HL CaC HC
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.
```

This command prints the Pearson correlation with two-tailed P-value and exclude the pair value if either GA or other independent variable has missing value.

		GA
Birth Weight	Pearson Correlation	.679**
	Sig. (2-tailed)	.000
	N	800
CHL	Pearson Correlation	.565**
	Sig. (2-tailed)	.000
	N	800
MAC	Pearson Correlation	.643**
	Sig. (2-tailed)	.000
	N	800
FL	Pearson Correlation	.432**
	Sig. (2-tailed)	.000
	N	800
HL	Pearson Correlation	.407**
	Sig. (2-tailed)	.000
	N	800
CaC	Pearson Correlation	.552**
	Sig. (2-tailed)	.000
	N	800
HC	Pearson Correlation	.517**
	Sig. (2-tailed)	.000
	N	800

The gestational age is significantly ($P < 0.001$) linearly correlated with all the seven independent variables. GA is highly correlated with BirthWeight and MAC. The strength of correlation of baby's anthropometry with gestational age varies from 0.407 to 0.679.

To further assess the collinearity among the independent variables, again Pearson correlation can be used.

Command to compute the Pearson correlation:

```
CORRELATIONS
/VARIABLES=BirthWeight CHL HC MAC FL HL CaC
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.
```

This command display the matrix with correlation coefficients, P-values (two-tailed), and number of pairs considered to calculate the correlation

Correlations

		Birth Weight	CHL	HC	MAC	FL	HL	CaC
Birth Weight	Pearson Correlation	1	.702**	.625**	.757**	.507**	.501**	.715**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	800	800	800	800	800	800	800
CHL	Pearson Correlation	.702**	1	.607**	.695**	.432**	.400**	.599**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000
	N	800	800	800	800	800	800	800
HC	Pearson Correlation	.625**	.607**	1	.607**	.439**	.333**	.528**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000
	N	800	800	800	800	800	800	800
MAC	Pearson Correlation	.757**	.695**	.607**	1	.500**	.458**	.720**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000
	N	800	800	800	800	800	800	800
FL	Pearson Correlation	.507**	.432**	.439**	.500**	1	.588**	.485**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000
	N	800	800	800	800	800	800	800
HL	Pearson Correlation	.501**	.400**	.333**	.458**	.588**	1	.434**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000
	N	800	800	800	800	800	800	800
CaC	Pearson Correlation	.715**	.599**	.528**	.720**	.485**	.434**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	
	N	800	800	800	800	800	800	800

**. Correlation is significant at the 0.01 level (2-tailed).

Some of the correlations among the independent variables are more than 0.7 and indicate that some variables have high collinearity. These should be automatically excluded by stepwise procedure. We use the method available in SPSS and not AIC we used in the book.

Command to run stepwise regression:

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA COLLIN TOL
```

```
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT GA
/METHOD=STEPWISE BirthWeight CHL HC MAC FL HL CaC
/RESIDUALS HIST(ZRESID).
```

This command runs stepwise multiple linear regression after excluding the cases for whom variable values are missing. The STATISTICS option provides the 95% CI on regression coefficients, and tests collinearity (by variance inflation factor (VIF)); CRITERIA option specifies that a variable be entered into the model if $P < 0.05$ and to remove it from the model is $P > 0.10$. Same criteria are used for all the variables in this example although you can specify different criteria for different variables. The RESIDUALS command is to draw a histogram and superimpose Gaussian curve to check the Gaussianity of error terms.

The following table displays the sequence of entering and removing of variables according to the criteria specified in the command. In this case no variable was removed.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Birth Weight	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	MAC	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	HC	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
4	FL	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: GA

Following table shows the model summary. The values of R-square show that BirthWeight is the first to enter and explained about 46.1% of variation in GA. Next is MAC that explained additional 3.9% variation. Then came HC and FL that explained 0.4% and 0.3% additional variance, respectively. All these are statistically significant. No other variable contributed significantly. Total variance of gestational age explained by four anthropometric variables is 50.7%. This is not high and the regression is not likely to provide accurate prediction.

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.679[a]	.461	.460	1.602
2	.707[b]	.500	.498	1.544
3	.710[c]	.504	.503	1.538
4	.712[d]	.507	.505	1.534

a. Predictors: (Constant), Birth Weight

b. Predictors: (Constant), Birth Weight, MAC

c. Predictors: (Constant), Birth Weight, MAC, HC

d. Predictors: (Constant), Birth Weight, MAC, HC, FL

e. Dependent Variable: GA

Following table depicts the analysis of variance that provides results of tests of the overall significance of the model at each stage of the stepwise procedure.

ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1749.125	1	1749.125	681.313	.000[a]
	Residual	2048.695	798	2.567		
	Total	3797.820	799			
2	Regression	1897.676	2	948.838	397.983	.000[b]
	Residual	1900.144	797	2.384		
	Total	3797.820	799			
3	Regression	1915.975	3	638.658	270.145	.000[c]
	Residual	1881.845	796	2.364		
	Total	3797.820	799			
4	Regression	1926.782	4	481.695	204.671	.000[d]
	Residual	1871.038	795	2.354		
	Total	3797.820	799			

a. Predictors: (Constant), Birth Weight

b. Predictors: (Constant), Birth Weight, MAC

c. Predictors: (Constant), Birth Weight, MAC, HC

d. Predictors: (Constant), Birth Weight, MAC, HC, FL

e. Dependent Variable: GA by Ballard

The following table shows steps in building the model with unstandardized or standardized coefficients of variables included in the model. Out of seven initial variables only four are entered into the model, others are not significant. All coefficients are positive that indicates these variables are positively associated with gestational age. The standardized coefficients are unitless and used to find the ranking of variable according to their importance. For example, in this case,

BirthWeight is more than six times as important as FL. Step four gives the final model equation.

The regression equation from the last block of the following table is
Gestational age = 24.734 + 1.508(birth weight) + 0.414(mid-arm circumference) + 0.080(head circumference) +0 .177(foot length)
This is different from the one obtained by R in the book.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics
		B	Std. Error	Beta			Lower Bound	Upper Bound	VIF
1	(Constant)	30.226	.246		122.681	.000	29.743	30.710	
	Birth Weight	2.571	.098	.679	26.102	.000	2.378	2.764	1.000
2	(Constant)	27.595	.409		67.431	.000	26.792	28.398	
	Birth Weight	1.702	.145	.449	11.711	.000	1.417	1.987	2.345
	MAC	.483	.061	.303	7.894	.000	.363	.603	2.345
3	(Constant)	25.390	.891		28.483	.000	23.640	27.139	
	Birth Weight	1.566	.153	.414	10.257	.000	1.267	1.866	2.611
	MAC	.437	.063	.274	6.923	.000	.313	.561	2.517
	HC	.090	.032	.092	2.782	.006	.026	.153	1.763
4	(Constant)	24.734	.941		26.295	.000	22.887	26.580	
	Birth Weight	1.508	.155	.398	9.744	.000	1.204	1.812	2.694
	MAC	.414	.064	.260	6.480	.000	.288	.539	2.590
	HC	.080	.033	.082	2.473	.014	.017	.144	1.795
	FL	.177	.083	.064	2.143	.032	.015	.340	1.431

a. Dependent Variable: GA

Residual statistics in the following table show that the standardized residuals range from -3.520 to +3.143 and the SD is nearly 1. This is acceptable. VIF = $1/(1 - R^2)$, where R is calculated with the remaining independents. Generally a value of VIF more than 5 is considered to indicate a mild and more than 10 a strong collinearity. In this case, none of the selected independents have high collinearity.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	31.18	41.29	36.49	1.553	800
Residual	-5.400	4.821	.000	1.530	800
Std. Predicted Value	-3.419	3.095	.000	1.000	800
Std. Residual	-3.520	3.143	.000	.997	800

a. Dependent Variable: GA

The histogram of standardized residuals, superimposed with Gaussian curve shows an approximate Gaussian distribution of error term (Figure C.4(b)).

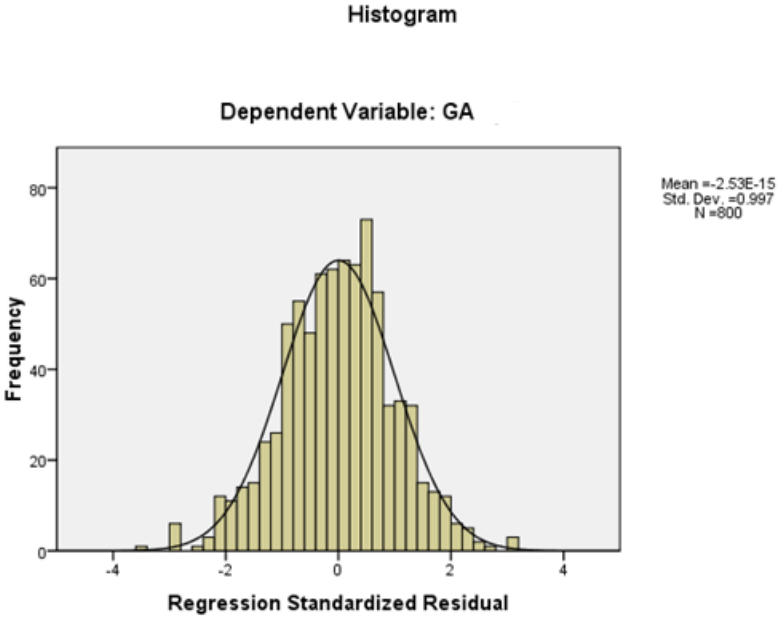

FIGURE C.4(b) Checking the validity of requirements of multiple regression