8. Selecting a Regression Model

In this section we will use some variable selection techniques to obtain a new regression model for predicting birth weight. These techniques include backward elimination procedure, forward regression, and stepwise regression. The best subsets method is not supported by SPSS. The techniques are based on adding independent variables (one at a time) to a regression model or removing independent variables (one at a time) from the model.

Forward variable selection enters the variables into the model one at a time based on entry criteria. At each step, the hypothesis that the coefficient of the entered variable is 0 is tested using its t statistic (actually an F statistic that is the square of the t). *Backward* variable elimination begins with all independent variables in the model, and at each step, removes the least useful predictor. Variables are removed until an established criterion holds. *Stepwise* selection begins like forward method, but at each step, tests variables already in the model for removal.

SPSS provides two criteria for moving variables. They are based on an F statistic that is the square of the t statistic. The first criterion for removing variables is the minimum F value that a variable must have to remain in the model. Variables with F statistics less than the value specified for removal are eligible for removal. Some texts and software packages call this statistic *F-to-remove*. The second criterion is the *maximum probability of F-to-remove*. The default F-to-remove is 2.71, and the default probability is 0.10.

Forward selection method applied our data can be summarized by the following table:

FORWARD SELECTION METHOD				
STEP	**VARIABLE ENTERED**	**VARIABLES IN THE MODEL**	**R^2**	**ADJ. R^2**
1	GESTWKS	GESTWKS	.18135	.18014
2	MPPWT	GESTWKS, MPPWT	.22128	.21898
3	MNOCIG	GESTWKS, MPPWT, MNOCIG	.24237	.23901
4	FHEIGHT	GESTWKS, MPPWT, MNOCIG, FHEIGHT	.25556	.25115
5	MHEIGHT	GESTWKS,MPPWT, MNOCIG, FHEIGHT, MHEIGHT	.26140	.25592

The criterion *Probability-of-F-to-enter <= .050* is used to enter a variable into the model.

The stepwise regression applied to our data can be summarized in the following table:

STEPWISE SELECTION METHOD				
MODEL	**VARIABLES ENTERED**	**VARIABLES REMOVED**	**R^2**	**ADJ. R^2**
1	GESTWKS		.18135	.18014
2	MPPWT		.22128	.21898
3	MNOCIG		.24237	.23901
4	FHEIGHT		.25556	.25115
5	MHEIGHT		.26140	.25592

As you can see, the final regression model produced by SPSS includes the same five variables found via the forward selection method.

The backward elimination method applied to our set of independent variables can be summarized in the following table:

BACKWARD ELIMINATION METHOD				
MODEL	**VARIABLES REMOVED**	**VARIABLES IN THE MODEL**	**R^2**	**ADJ. R^2**
1		GESTWKS,MNOCIG,MAGE, MHEIGHT,MPPWT,FEDYRS, FHEIGHT, FNOCIG, FAGE	.26194	.25203
2	FAGE	GESTWKS,MNOCIG,MAGE, MHEIGHT,MPPWT,FEDYRS, FHEIGHT, FNOCIG	.26194	.25314
3	MAGE	GESTWKS,MNOCIG,MHEIGHT, MPPWT,FEDYRS,FHEIGHT, FNOCIG	.26191	.25422
4	FEDYRS	GESTWKS,MNOCIG,MHEIGHT, MPPWT, FHEIGHT, FNOCIG	.26186	.25528
5	FNOCIG	GESTWKS,MNOCIG,MPPWT, FHEIGHT, MHEIGHT	.26140	.25592

The backward elimination method has produced the same regression model with the five independent variables GESTWKS, MNOCIG, MPPWT, FHEIGHT, and MHEIGHT.

The regression model obtained via the three variable selection techniques has the form:

$$BWT = \beta_0 + \beta_1 * GESTWKS + \beta_2 * MNOCIG + \beta_3 * MHEIGHT + \beta_4 * MPPWT +$$
$$+ \beta_5 * FHEIGHT + ERROR.$$

The SPSS output for the model is displayed below:

MULTIPLE LINEAR REGRESSION	
Multiple R	.51127
R Square	.26140
Adjusted R Square	.25592
Standard Error	.94226

Analysis of Variance

	DF	Sum of Squares	Mean Square
Regression	5	211.78250	42.35650
Residual	674	598.41303	.88785

F = 47.70665 Signif F = .0000

The squared multiple correlation coefficient R^2 (0.26140) is now slightly smaller than the value obtained for the full model (0.26194). The linear regression of birth weight on the five predictors explains almost the same portion (over 26 %) of the variation in infant birth weights as the model with the nine variables.

The sum of squares due to the regression model 211.78250 is slightly smaller than the value of 212.22479 obtained for the full model. The sum of squares due to error (residual sum of squares) 598.41303 is slightly larger than the value of 597.97074 obtained for the model with the nine variables. The residual mean square is an estimate of the variance σ^2 and is equal to 0.88785. The value of the F statistic equal to 47.70665 is now much larger than the value of 26.42095 obtained for the full model.

Now we analyze the part of the output providing the estimates of the regression parameters.

---------- **Variables in the Equation** ----------

Variable	B	SE B	Beta	VIF	T	Sig T
GESTWKS	.233851	.019363	.401496	1.008	12.077	.0000
MNOCIG	-.014619	.003220	-.150857	1.008	-4.540	.0000
MHEIGHT	.040042	.017358	.091028	1.421	2.307	.0214
MPPWT	.008308	.002330	.135979	1.327	3.565	.0004
FHEIGHT	.039711	.014387	.095913	1.102	2.760	.0059
(Constant)	-8.114024	1.40914			-5.758	.0000

According to the above output, the estimated regression line of birth weights on the five predictors is

$$\mu\{BWT\} = .234 * GESTWKS - .015 * MNOCIG + .040 * MHEIGHT + .008 * MPPWT + .040 * FHEIGHT - 8.114.$$

All independent variables in the model are significant with the p-value <= .0214. The regression diagnostics for the model is discussed in **Section 9**.