Chapter 4

Statistical Modeling

In Chapter 3, the preliminary data analysis presented the general characteristics of graduate students and the associations between academic achievement and each determinant. We found that faculty, BA GPA, and total years of study are statistically significantly associated with academic achievement. In this chapter multiple regression analysis is used for developing a predictive model for academic achievement. The total number of determinants is 15 (see Chapter 1) and the outcome is the academic achievement of graduate students, measured in terms of their grade point average. These measurements were based on a sample of 294 subjects. In this study, the main determinants are faculty, BA GPA, and total years of study.

Figure 4.1 shows the full model of the multiple regression analysis with all the predictor variables included. The model gives the p-value 0.000 for testing the null hypothesis of no linear relation between the determinants and the outcome. The goodness-of-fit, which is measured by the r-squared statistic, is 0.292. This means that this model can account for 29.2% of the variation in the outcomes. The standard deviation of the residual MA GPA is 0.22.

Next, for selecting the best regression equation, a stepwise procedure was used and using SPSS version 7.5 for windows and Matlab program and Asp for combine the predictors which have approximately the same of coefficients.

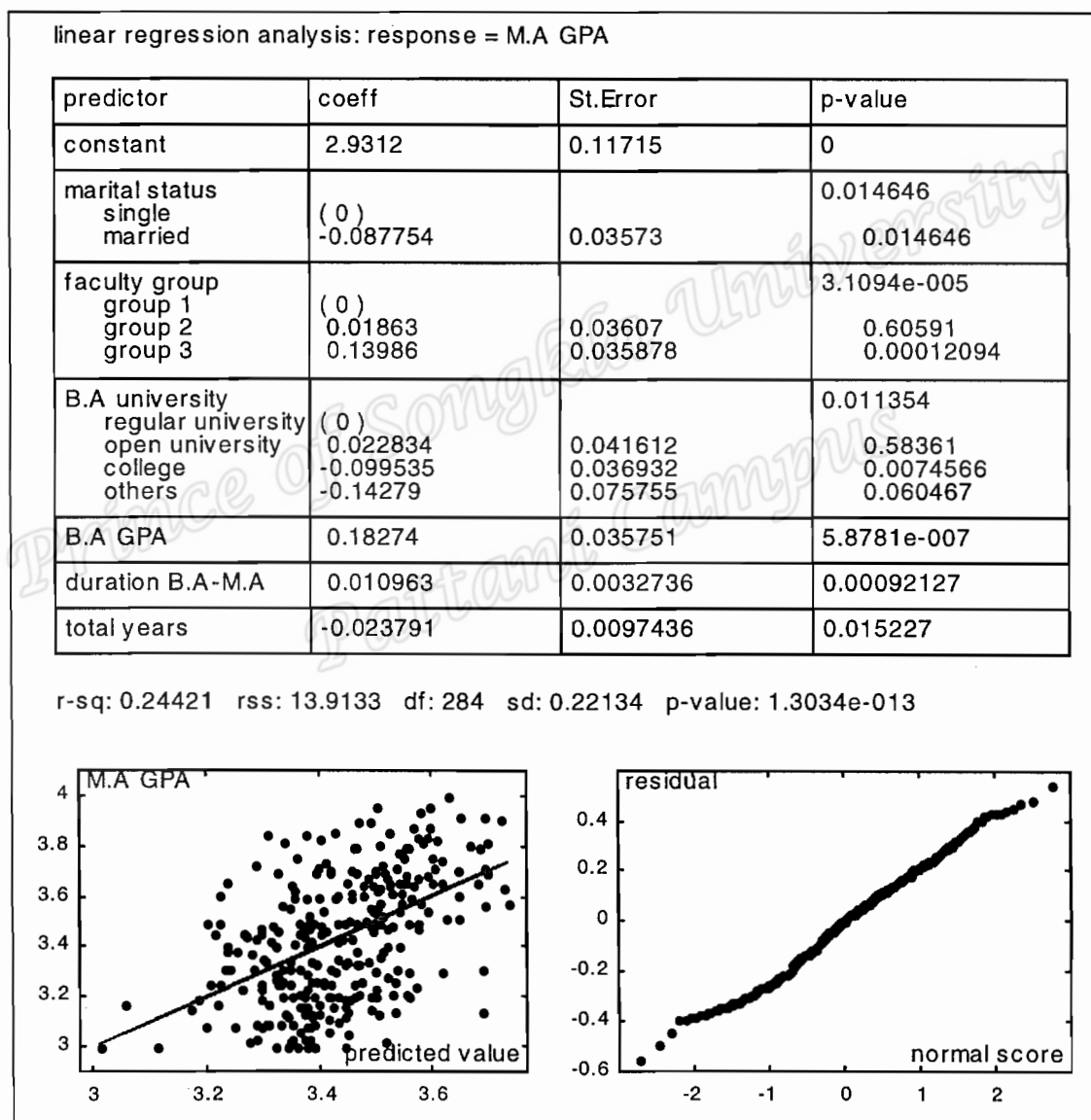*Figure 4.1 Full model of multiple regression analysis*

linear regression analysis: response = M.A GPA

predictor	coeff	St.Error	p-value
constant	2.7457	0.17801	0
gender 　male 　female	 (0) -0.022465	 0.029756	0.45092 0.45092
age	0.0022255	0.0052615	0.67265
marital status 　single 　married	 (0) -0.08528	 0.04176	0.042102 0.042102
domicile 　near 　local 　others	 (0) 0.027289 0.049405	 0.032242 0.044152	0.45897 0.39808 0.26414
occupation 　no work 　government 　individual	 (0) 0.061905 0.1177	 0.042406 0.048454	0.050912 0.1455 0.015781
experience	0.0013429	0.0054853	0.80678
faculty group 　group 1 　group 2 　group 3	 (0) 0.02072 0.15482	 0.039889 0.037607	4.2864e-005 0.60388 5.0965e-005
Study plan 　A 　B	 (0) -0.032054	 0.071152	0.65271 0.65271
Study Type 　full time 　part time	 (0) -0.048258	 0.076372	0.528 0.528
B.A University 　regular University 　open University 　college 　others	 (0) 0.066035 -0.1036 -0.16527	 0.047924 0.038582 0.081703	0.0013829 0.16936 0.007692 0.044073
B.A major 　science 　humanities 　education	 (0) -0.042754 0.020493	 0.038004 0.053473	0.40464 0.26159 0.70184
B.A program 　2 year 　4 year	 (0) 0.0776	 0.036637	0.035077 0.035077
B.A GPA	0.19166	0.036208	2.4765e-007
durationB.A-M.A	0.0085954	0.0045987	0.062683
total years	-0.024348	0.010337	0.019207

r-sq: 0.29215   rss: 13.0307   df: 272   sd: 0.21888   p-value: 9.1841e-012

The reduced model comprises six predictors (marital status, faculty, BA GPA, BA university, duration BA to MA, and total years of study). The percentage of the variation in the outcome accounted for by these predictors was 24.4%, with the residual standard deviation 0.22, the same value as for the full model. The result of fitting this reduced model is shown in Figure 4.2.

Figure 4.2 Reduced model of multiple regression analysis, with residuals

linear regression analysis: response = M.A GPA

predictor	coeff	St.Error	p-value
constant	2.9312	0.11715	0
marital status 　single 　married	 (0) -0.087754	 0.03573	0.014646 0.014646
faculty group 　group 1 　group 2 　group 3	 (0) 0.01863 0.13986	 0.03607 0.035878	3.1094e-005 0.60591 0.00012094
B.A university 　regular university 　open university 　college 　others	 (0) 0.022834 -0.099535 -0.14279	 0.041612 0.036932 0.075755	0.011354 0.58361 0.0074566 0.060467
B.A GPA	0.18274	0.035751	5.8781e-007
duration B.A-M.A	0.010963	0.0032736	0.00092127
total years	-0.023791	0.0097436	0.015227

r-sq: 0.24421 rss: 13.9133 df: 284 sd: 0.22134 p-value: 1.3034e-013

The normal scores plot in the bottom right of Figure 4.2 shows an approximately linear trend, suggesting that the normality assumption is reasonable for these data. However, recall that Figure 3.1 in Chapter 3 showed that the distributions of duration BA to MA and total years of study were skewed (1.075 and 1.778, respectively). For this reason it might be better to transform these variables. Both duration variables (duration from BA to MA and total years of study) are transformed by base 2 logarithms, so that each unit for the transformed data corresponds to a doubling of the time duration. However, since the duration from BA to MA can be 0, 1 is added before taking logarithms in this case. The distributions of data after this trans-formation are shown in Figure 4.3. The resulting regression is shown in Figure 4.4.

Figure 4.3 Distributions of remaining predictors (with duration log-transformed)

	col	variable	size	mean	st dev	SE	skew	kurt	min	max
	1	marital status	294	1.34	0.475	0.028	0.674	1.451	1	2
	2	faculty group	294	2.201	0.791	0.046	-0.372	1.692	1	3
	3	B.A university	294	1.554	0.872	0.051	1.273	3.27	1	4
	4	B.A GPA	294	2.794	0.398	0.023	0.452	2.635	2	3.93
	5	log2(duration+1)	294	2.242	1.221	0.071	-0.109	2.058	0	4.585
	6	log2(total year)	294	1.758	0.491	0.029	0.611	3.456	0.604	3.342
	7	MA GPA	294	3.437	0.251	0.015	0.105	2.015	3	4

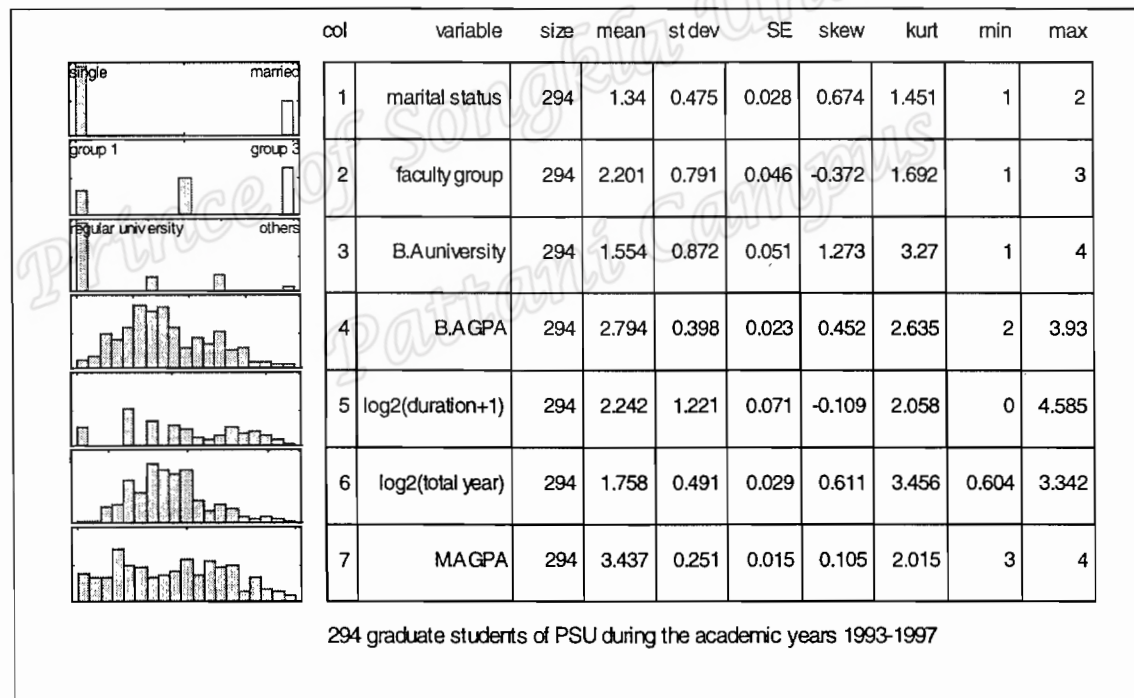294 graduate students of PSU during the academic years 1993-1997

Figure 4.3 clearly show that the distribution of base 2 logarithms of duration from BA to MA and total years of study are approximately normally distributed. The skewnesses of these two variables in this figure are lower (−0.109 and 0.611 respectively) than they are in Figure 3.1 of Chapter 3.

Figure 4.4 shows the resulting regression after transforming the durations. This model again has statistically highly significant predictors. The r-squared is higher (r-square = 26.3%) than for the previous model (shown as Figure 4.2). The regression coefficients for marital status, BA university (college and others), and total years of study are negative. The coefficients for faculty groups 1 and 2 are approximately the same, and also the coefficients for regular and open university are approximately the same, when college and others are nearly the same. Thus these four groups could be combined into two groups. Figure 4.5 is thus obtained.

Figure 4.4: Reduced model of multiple regression analysis, with residuals

linear regression analysis: response = M.A GPA

predictor	coeff	St.Error	p-value
constant	2.9068	0.12583	0
marital status			0.005414
single	(0)		
married	-0.095271	0.033991	0.005414
faculty group			8.9975e-005
group 1	(0)		
group 2	0.015115	0.035688	0.67222
group 3	0.12994	0.03558	0.00030958
B.A university			0.0084505
regular university	(0)		
open university	0.0094744	0.040802	0.81654
college	-0.10589	0.036348	0.0038621
others	-0.14371	0.0748	0.055705
B.A GPA	0.18964	0.035236	1.5429e-007
log2(duration+1)	0.055837	0.013461	4.4343e-005
log2(total year)	-0.076026	0.027472	0.0060212

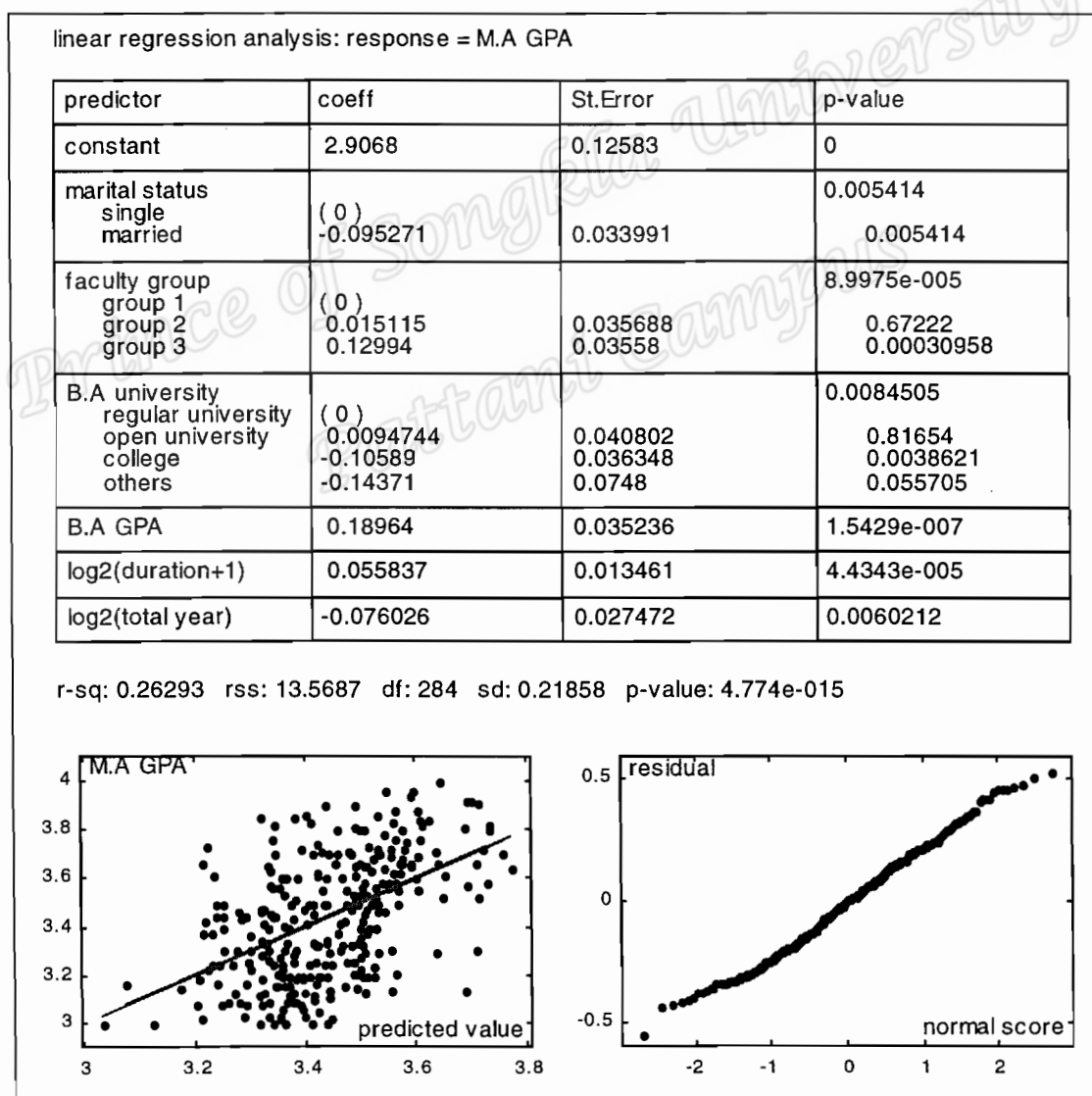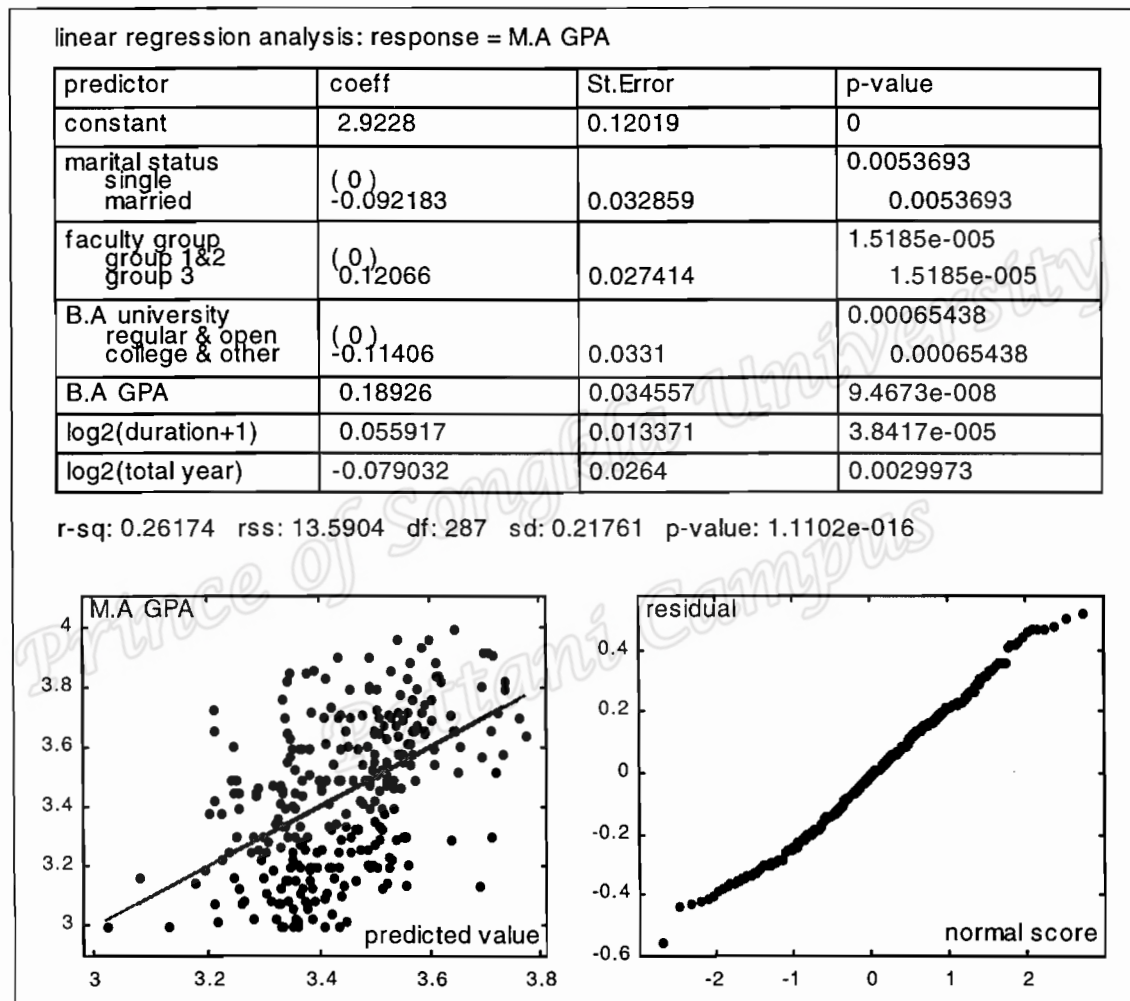r-sq: 0.26293   rss: 13.5687   df: 284   sd: 0.21858   p-value: 4.774e-015

Figure 4.5 shows the goodness-of-fit is 26.2%, which is the same as for the model that did not combine faculty groups 1 and 2, and the four groups of BA university. The regression coefficients for each of the predictors are unchanged.

Figure 4.5: Model with combined faculty groups and BA university

linear regression analysis: response = M.A GPA

predictor	coeff	St.Error	p-value
constant	2.9228	0.12019	0
marital status single married	(0) -0.092183	 0.032859	0.0053693 0.0053693
faculty group group 1&2 group 3	(0) 0.12066	 0.027414	1.5185e-005 1.5185e-005
B.A university regular & open college & other	(0) -0.11406	 0.0331	0.00065438 0.00065438
B.A GPA	0.18926	0.034557	9.4673e-008
log2(duration+1)	0.055917	0.013371	3.8417e-005
log2(total year)	-0.079032	0.0264	0.0029973

r-sq: 0.26174   rss: 13.5904   df: 287   sd: 0.21761   p-value: 1.1102e-016

We now consider the interaction effect in the multiple regression analysis model. We focus on BA GPA and duration from BA to MA, because these variables are highly statistically significant. There might be an interaction between them. So fitting the model with an interaction term between the BA GPA and duration BA to MA, and the result is shown in Figure 4.6.

Figure 4.6: Model of multiple regression analysis with interaction

linear regression analysis: response = MAGPA

predictor	coeff	St.Error	p-value
constant	2.7458	0.21037	0

marital status
  single
  married
	(0)
	-0.09614
		
		0.033083
			0.0039463
			
			0.0039463

faculty group
  group 1&2
  group 3
	(0)
	0.12307
		
		0.027513
			1.1124e-005
			
			1.1124e-005

B.A university
  regular & open
  college & other
	(0)
	-0.11552
		
		0.033128
			0.00056483
			
			0.00056483

B.AGPA	0.25051	0.069032	0.00033701
log2(duration+1)	0.13748	0.080693	0.089517
log2(total year)	-0.078146	0.026412	0.0033475
B.AGPA*log2(duration+1)	-0.02906	0.028353	0.30625

r-sq: 0.26444 rss: 13.5407 df: 286 sd: 0.21759 p-value: 2.2204e-016

After fitting the model with an interaction term, a p-value of interaction term (BA and duration BA to MA) had the lowest the p-value, as shown in the bottom panel of Figure 4.6. The interaction is thus not significant, and the r-squared and regression coefficients of each predictor are the same. We conclude that the interaction effect is not needed. Thus the model in Figure 4.5 is sufficient to describe the data. It can account for 26.2% of the variation in academic achievement. BA GPA had a slightly highest effect (regression coefficients 18.9%). The model takes the form

$$
\begin{aligned}
\textit{Academic Achievement} \ = \ & 2.923 - 0.092 \ \textit{married} + 0.121 \ \textit{faculty group 3} \\
& - 0.114 \ \textit{college \& others university} + 0.189 \ \textit{BA GPA} \\
& + 0.056 \ \log_2(\textit{duration BA to MA+1}) \\
& - 0.079 \ \log_2(\textit{total years of study}).
\end{aligned}
$$

Where, Faculty group 3 comprise Nursing, Education, Humanities, Natural Resources, and faculty group 2 comprise Agro-Industrial, Engineering, Management Science, and Science &Technology. The other two faculties (Environment Management, Science) belong to group 1.