

# Applied Linear Model 1 - STA 5002

HUONG TRAN

<sup>a</sup>*Oakland University*

---

**The Problem:** In the early of 21st century, colleges in Michigan faced big problem with keeping student involved. To be more specific, although they could attract new student, but getting student to graduate was a difficult task. From Integrated Postsecondary Education Data System (IPEDS), the percentage of student who registered in 2003 and graduated in 2009, is less than 50% in many universities in Michigan. Numerous reasons including academic preparation, tuition fees or personal finance were considered as main factors causing the low rate of graduation. This regression analysis aims to assess the impact of these factors and estimate a model that may help to predict the graduation rate of a university.

**The Data:** The data was collected from 15 universities in Michigan ( $n = 15$  observations) in 2010, and the data base is available on the website of Integrated Postsecondary Education Data System (IPEDS). The dependent variable of interest is  $y$ , the percentage of graduation who were first-time, full-time college students in 2003 and had earned a degree within 6 years (they graduated in 2009).

The independent variables are described as follow:

- Enrollment: Total number of students enrolled to the college.
- Tuition for 30 credits (dollars): Tuition and mandatory fees for a freshmen taking 30 credit hours.
- Average GPA: The GPA for 2010-year's freshmen class. This help to measure the academic preparation of new students, which is good to get intuition about the quality of students before they register to a university.
- Average ACT: The ACT for 2010-year's freshmen class. This help to measure the academic preparation of new students, which is good to get intuition about the quality of students before they registry.
- Retention Rate: Percentage of students who began in Fall 2008 and returned in Fall 2009. This may help to measure the involvement between freshmen and the university.
- Full- time : The percentage of full-time student in college.

---

*Email address:* [nguyenquynhhuon@oakland.edu](mailto:nguyenquynhhuon@oakland.edu) (HUONG TRAN)

- Percent undergrads over 25: The percentage of undergraduate students who are over 25 years old.

**Variable Screening:** In this step, we will find the "important" independent variables which contribute most to the prediction of percentage of graduation.

- Forward selection method: Figure 1 illustrate the forward selection screening method, from which, the two most important variables are the percentage of retention and the percentage of full time student.

	Coef	P	Coef	P
Constant	-92.5		-102.2	
Ret Rate (x5)	1.864	0.000	1.529	0.000
Full Time (x6)			0.436	0.064
S		7.73657		6.93514
R-sq		80.31%		85.39%
R-sq(adj)		78.79%		82.96%
Mallows' Cp		14.06		9.59
AICc		109.98		109.32
BIC		109.92		108.15

*a to enter = 0.15*

Figure 1: Forward selection method

- As discuss above, GPA and ACT are represented for the academic preparation of student, i.e, the quality of student entering the university. These two variables are important. However, looking at their coefficient of correlation, in figure 2, which is 0.931. Because the two variables are highly correlated, it may cause multicollinearity in our model. To cope with this issue, it is recommended to eliminate one variables and keep one in our model. Therefore, I will choose GPA as a predictor of percentage of graduation rate.

MI GRAD RATES.MTW	
<b>Correlation: GPA (x3), ACT (x4)</b>	
<b>Correlations</b>	
	<b>GPA (x3)</b>
ACT (x4)	0.931

Figure 2: Coefficient of correlation of  $x_3$  and  $x_4$

In conclusion, this analysis and intuition leads us to select the following variables to begin model-building process:

- Average GPA:  $x_1$

- Retention rate:  $x_2$
- Percentage of full-time student:  $x_3$

**Intuition about the relationship of each predictors to the percentage of graduation:**

The plot of percentage of graduation ( $y$ ) versus GPA, retention rate and percentage of full-time student is in figure 3.

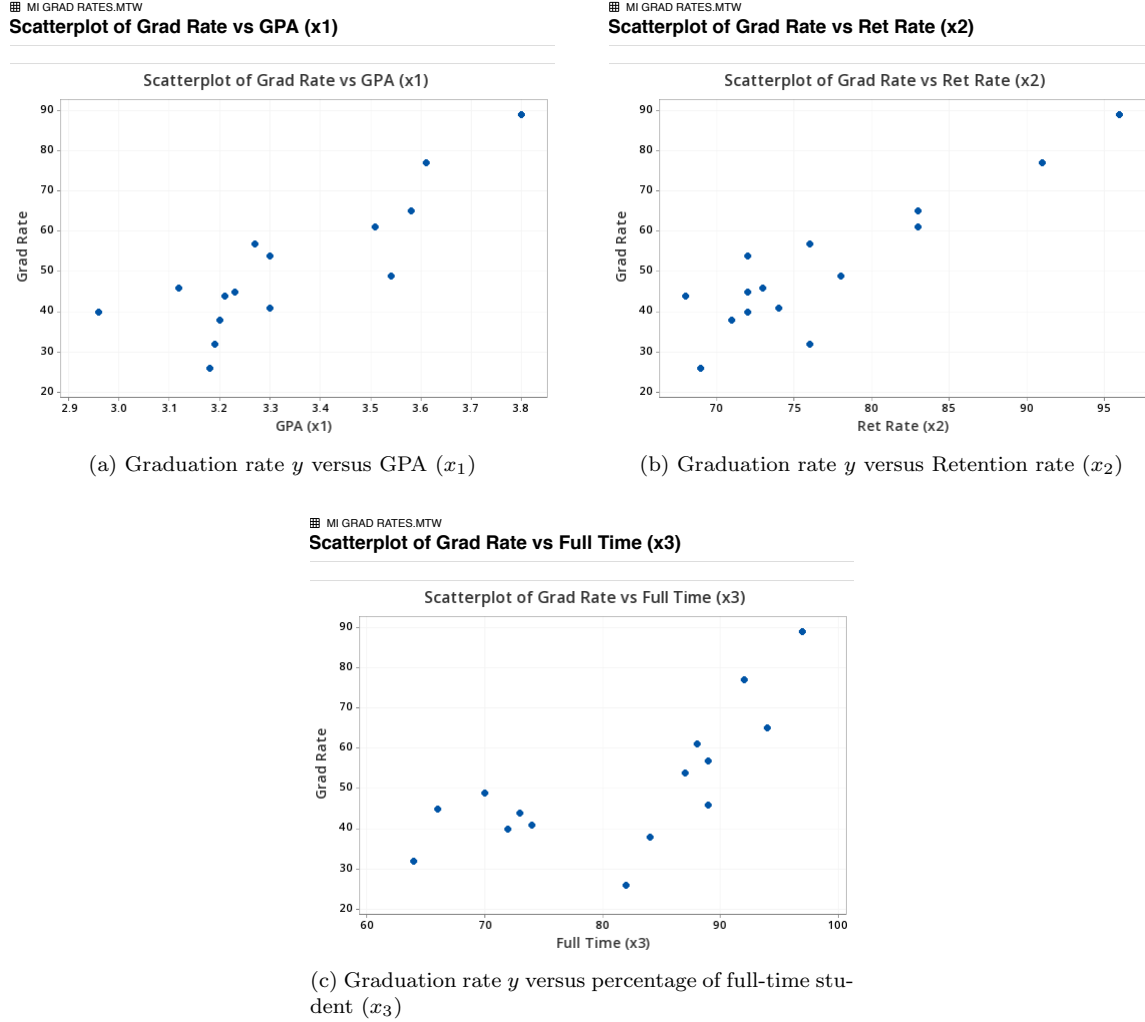


Figure 3: Scatter plot about the relationship of each predictions to the percentage of graduation ( $y$ )

**Coded Variables:** The range of Average GPA  $x_1$  is 0.84 while the range of Retention rate and percentage of full-time student are 28 and 33 respectively. Therefore, the model may get the result

with large rounding errors caused by the sizable disparity in the ranges of predictors. To avoid that, we should code the independent variables, and the procedure is as follow:

$$u_i = \frac{x_i - \bar{x}}{s_x}, \text{ where } s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Coded variables  $u_1, u_2, u_3$  are stored in figure 4.


	C1-T	C2	C3	C4	C5 	C6	C7	C8	C9
	Inst	GPA (x1)	Ret Rate (x2)	Full Time (x3)	Grad Rate	u1	u2	u3	
2	EMU	2.96	72	72	40	-1.66022	-0.61092	-0.88050	
3	FSU	3.21	68	73	44	-0.54846	-1.10626	-0.78683	
4	GVSU	3.51	83	88	61	0.78564	0.75127	0.61822	
5	LSSU	3.18	69	82	26	-0.68187	-0.98243	0.05620	
6	MSU	3.61	91	92	77	1.23034	1.74195	0.99291	
7	MTU	3.58	83	94	65	1.09693	0.75127	1.18025	
8	NMU	3.12	73	89	46	-0.94869	-0.48709	0.71189	
9	OU	3.30	74	74	41	-0.14823	-0.36325	-0.69316	
10	SVSU	3.20	71	84	38	-0.59293	-0.73476	0.24354	
11	UMAA	3.80	96	97	89	2.07527	2.36113	1.46126	
12	UMD	3.54	78	70	49	0.91905	0.13209	-1.06784	
13	UMF	3.23	72	66	45	-0.45952	-0.61092	-1.44252	
14	WSU	3.19	76	64	32	-0.63740	-0.11558	-1.62986	
15	WMU	3.30	72	87	54	-0.14823	-0.61092	0.52455	
16									

Figure 4: Coded variables  $u_1, u_2, u_3$

**Model Building:** Now, we will hypothesize the model for prediction of percentage of graduation as follow:

1. Model 1: Complete second-order model:

$$\begin{aligned} E(y) = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ & + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 \\ & + \beta_7 x_1^2 + \beta_8 x_2^2 + \beta_9 x_3^2 \end{aligned}$$

2. Model 2: Interaction model:

$$\begin{aligned} E(y) = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ & + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 \end{aligned}$$

3. Model 3: First-order model:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

**Test for significance of All Quadratic terms (Model 1 and Model 2):**

$$H_0 : \beta_7 = \beta_8 = \beta_9 = 0$$

$H_a$  : At least one the quadratic  $\beta$ 's in Model 1 differs from 0

Minitab printout the Complete second-order model for the percentage of graduation from colleges in 2010 in figure 5.

## Regression Analysis: Grad Rate versus u1, u2, u3

### Regression Equation

$$\text{Grad Rate} = 45.12 + 0.96 u1 + 5.36 u2 + 11.55 u3 + 14.11 u1*u1 + 15.0 u2*u2 + 3.29 u3*u3 - 38.6 u1*u2 - 6.62 u1*u3 + 19.65 u2*u3$$

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	45.12	5.27	8.55	0.000	
u1	0.96	5.05	0.19	0.857	7.90
u2	5.36	6.10	0.88	0.420	11.53
u3	11.55	4.33	2.67	0.044	5.80
u1*u1	14.11	9.17	1.54	0.185	36.04
u2*u2	15.0	14.0	1.07	0.335	134.93
u3*u3	3.29	4.45	0.74	0.492	3.96
u1*u2	-38.6	24.6	-1.57	0.177	294.00
u1*u3	-6.62	5.37	-1.23	0.272	8.99
u2*u3	19.65	9.04	2.17	0.082	24.78

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
6.72321	94.28%	83.98%	0.00%

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	9	3724.93	413.881	9.16	0.013
u1	1	1.62	1.624	0.04	0.857
u2	1	34.83	34.826	0.77	0.420
u3	1	321.88	321.879	7.12	0.044
u1*u1	1	107.00	107.000	2.37	0.185
u2*u2	1	51.49	51.495	1.14	0.335
u3*u3	1	24.82	24.815	0.55	0.492
u1*u2	1	111.26	111.258	2.46	0.177
u1*u3	1	68.70	68.704	1.52	0.272
u2*u3	1	213.40	213.403	4.72	0.082
Error	5	226.01	45.202		
Total	14	3950.93			

Figure 5: Complete second-order model for predicting  $y$

Minitab printout the Interaction model in figure 6

## Regression Analysis: Grad Rate versus u1, u2, u3

### Regression Equation

Grad Rate = 50.43 - 0.62 u1 + 9.18 u2 + 7.06 u3 - 6.75 u1\*u2 - 4.67 u1\*u3 + 15.50 u2\*u3

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	50.43	2.58	19.58	0.000	
u1	-0.62	4.84	-0.13	0.901	7.51
u2	9.18	5.29	1.74	0.121	8.95
u3	7.06	2.65	2.66	0.029	2.25
u1*u2	-6.75	4.70	-1.44	0.189	11.06
u1*u3	-4.67	4.67	-1.00	0.347	7.01
u2*u3	15.50	7.87	1.97	0.084	19.37

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
6.61315	91.14%	84.50%	62.86%

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	6	3601.06	600.177	13.72	0.001
u1	1	0.72	0.723	0.02	0.901
u2	1	131.68	131.681	3.01	0.121
u3	1	310.43	310.435	7.10	0.029
u1*u2	1	90.13	90.134	2.06	0.189
u1*u3	1	43.70	43.704	1.00	0.347
u2*u3	1	169.81	169.809	3.88	0.084
Error	8	349.87	43.734		
Total	14	3950.93			

Figure 6: Interaction model for predicting  $y$

- Test statistic:

$$\begin{aligned}
 F &= \frac{(SSE_R - SSE_C)/(k - g)}{MSE_C} \\
 &= \frac{(349.87 - 226.01)/3}{45.202} \\
 &= 0.9134
 \end{aligned}$$

- Degree of freedom for the numerator is number of  $\beta$  tested: 3
- Degree of freedom for denominator is:  $n - (k + 1) = 15 - 10 = 5$ .

## Probability Density Function

F distribution with 3 DF in numerator and 5 DF in denominator

x	f(x)
0.9134	0.393912

Figure 7:  $p$ -value

Therefore,  $p\text{-value} = 0.393912 > \alpha = 0.05$  (result from Minitab in 7), we fail to reject the null hypothesis. Therefore, the data does not provide sufficient evidence of second-order terms, we prefer the interaction model, i.e, Model 2.

**Test for significance of Interaction terms (Model 2 and Model 3):**

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$$

$H_a$  : At least one the interaction coefficient  $\beta$ 's in Model 2 differs from 0

Information about interaction model are available in figure 6. Minitab printout the First-order model in figure 8.



## Regression Analysis: Grad Rate versus u1, u2, u3

### Regression Equation

$$\text{Grad Rate} = 50.93 + 3.38 u1 + 9.48 u2 + 4.49 u3$$

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	50.93	1.81	28.13	0.000	
u1	3.38	3.94	0.86	0.409	4.41
u2	9.48	4.06	2.33	0.040	4.69
u3	4.49	2.31	1.94	0.078	1.52

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
7.01272	86.31%	82.57%	75.17%

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	3409.97	1136.66	23.11	0.000
u1	1	36.19	36.19	0.74	0.409
u2	1	268.11	268.11	5.45	0.040
u3	1	185.16	185.16	3.77	0.078
Error	11	540.96	49.18		
Total	14	3950.93			

Figure 8: First-order model for predicting  $y$

- Test statistic:

$$\begin{aligned}
 F &= \frac{(SSE_R - SSE_C)/(k - g)}{MSE_C} \\
 &= \frac{(540.96 - 349.87)/3}{43.734} \\
 &= 1.4565
 \end{aligned}$$

- Degree of freedom for the numerator is number of  $\beta$  tested: 3
- Degree of freedom for denominator is:  $n - (k + 1) = 15 - 7 = 8$ .

## Probability Density Function

### F distribution with 3 DF in numerator and 8 DF in denominator

x	f(x)
1.4565	0.248268

Figure 9:  $p$ -value

Therefore,  $p\text{-value} = 0.2482682 > \alpha = 0.05$  (result from Minitab in 9), we fail to reject the null hypothesis. Therefore, the data does not provide sufficient evidence of interaction terms, we prefer the first-order model, i.e, Model 1.

**Conclusion:** First-order model outweighs the interaction model and the second-order model. Looking at the model in figure 8 more details:

- In this model,  $F\text{-value} = 23.11$  and  $p\text{-value} = 0.000 < \alpha = 0.05$ , we reject the null hypothesis  $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$  and conclude that this model is useful to predict the percentage of graduation.
- $R^2 = 86.31$ : there are 86.31% of mean value of percentage of students graduating within 6 years can be explained by this model.
- However, making prediction outside observational data will be a dangerous task. Therefore, it is recommend to make predict with data in experiment region, GPA -  $x_1 \in [2.96, 3.8]$ , retention rate -  $x_2 \in [68, 96]$  and percentage of full time students -  $x_3 \in [64, 97]$ .

From figure 8, the model is:

$$E(\hat{y}) = 50.93 + 3.38u_1 + 9.48u_2 + 4.49u_3$$

From the model, we can have some comments:

- $\beta_1 = 3.38$ : if the average GPA of new student increase 0.1 point, there will be 0.34 percent of students graduating within 6 years more, provided retention rate and percentage of full-time student are fixed.
- $\beta_2 = 9.48$ : if the percentage of retention increase 1 percent, there will be 9.48 percent of students graduating within 6 years more, provided GPA and percentage of full-time students are fixed.
- $\beta_3 = 4.49$ : if the percentage of full-time students increase 1 percent, there will be 4.5 percent of students graduating within 6 years more, provide the average GPA and retention rate are fixed.

**Recommendations:** The percentage of students graduating within 6 years will increase as average GPA or retention rate or percentage of full-time student increase. Below are some recommendations to improve graduation rate:

- Enhance the academic preparation for students in high-school. It is obvious to say that good foundation help student find interest in studying and keep them involved with the curriculum while students who are not carefully prepared may struggle with demanding subjects in university and get bored quickly.
- Large percentage of part-time students may lead to the low rate of graduation. For those part-time students, school may not be the first priority. In other words, they may have many other things to worry about: kids, financial burden, etc. It is a good idea if university can provide part-time jobs for student, so that they can overcome the finance problem and continue their study. Moreover, this is also help to get student more interact with the university.
- The interaction between new students and the university is the main factor to keep student return. Aggressive support should be considered as a priority to help student overcome obstacles of the first year. While GTAs (graduate teaching assistant) may help them figure out new knowledge, academic advisors guide them to a specific plan to finish the curriculum. In addition, other outside activities should be held to establish friendship and communication within students as well as campus organizers, from that, new students will get involved more and feel connected to campus.