STA 674

Regression Analysis And Design Of Experiments

Fitting Multiple Linear Regression Models – Lecture 7

STA 674, RADOE:
Fitting Multiple Linear Regression Models

- Last time, we talked about indicator variables..
- This time, we are going to look at using more than one indicator variable (a possibility, but usually more complex than necessary for a non-statistician) and consider what to do if the slopes in the two subgroups *aren't* similar.

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: US Graduation Rates

- Simple Linear Regression – Private only

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.85220	0.03543	24.06	<.0001	0.78190	0.92250
ADMISRATE	1	-0.30527	0.07611	-4.01	0.0001	-0.45630	-0.15423

- Simple Linear Regression – Public only

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.63554	0.05988	10.61	<.0001	0.51663	0.75446
ADMISRATE	1	-0.42084	0.08898	-4.73	<.0001	-0.59754	-0.24414

Interpretation: For every 1% increase in admission rate, private schools lost ~0.3% graduation rate, public schools lost ~0.42% graduation rate

MLR using indicator variable had a significant effect on graduation rate (0.28% difference, P=<0.001, t=11.75). For any given admission rate, private schools had a 0.28% higher graduation rate than a public school.

STA 674, RADOE:

Fitting Multiple Linear Regression Models

Example: US Graduation Rates

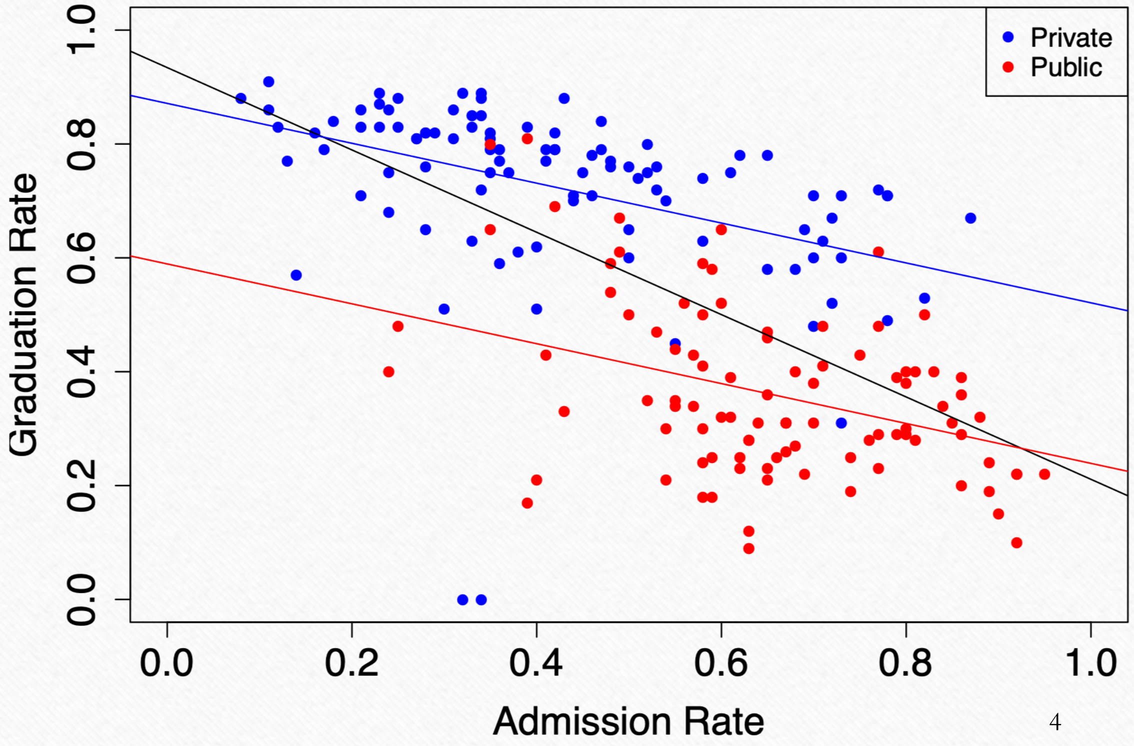
- Separate Intercept Model –

  Private versus Public

Root MSE	0.13975	R-Square	0.6551
Dependent Mean	0.54585	Adj R-Sq	0.6515
Coeff Var	25.60326		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	0.58944	0.04034	14.61	<.0001	0.50988	0.66901
ADMISRATE	1	-0.35044	0.05759	-6.09	<.0001	-0.46403	-0.23685
privateind	1	0.28196	0.02399	11.75	<.0001	0.23464	0.32928



STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: US Graduation Rates

```
DATA COLLEGE;
SET COLLEGE;

IF ivy="Yes" THEN ivyind=1;
ELSE ivyind=0;
RUN;
```

Obs	SCHOOL	GRADRATE4	ADMISRATE	PRIVATE	privateind	IVY	ivyind
1	Amherst College	0.84	0.18	Yes	1	No	0
2	Appalachian State University	0.31	0.64	No	0	No	0
3	Auburn University	0.4	0.83	No	0	No	0
4	Babson College	0.77	0.48	Yes	1	No	0
5	Bard College	0.59	0.36	Yes	1	No	0

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: US Graduation Rates

- Predictors

  x_1 = admission rate (ADMISRATE)

  x_2 = 0 for public schools, 1 for private schools (PRIVATEIND)

- Complete Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Model for Public Schools ($x_2 = 0$)

$$y = \beta_0 + \beta_1 x_1$$

- Model for Private Schools($x_2 = 1$)

$$y = (\beta_0 + \beta_2) + \beta_1 x_1$$

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: US Graduation Rates

- Predictors

  x_1 = admission rate (ADMISRATE)

  x_2 = 0 for public schools, 1 for private schools (PRIVATEIND)

  x_3 = 0 for non Ivy League schools, 1 for Ivy League schools (IVYIND) additional indicator added here = x3

- Complete Model

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

- Model for Public Schools ($x_2 = 0$ and $x_3 = 0$)

$$y = \beta_0 + \beta_1 x_1$$

- Model for Private, non Ivy League Schools ($x_2 = 1$ and $x_3 = 0$)

$$y = (\beta_0 + \beta_2) + \beta_1 x_1$$

- Model for Ivy League Schools ($x_2 = 1$ and $x_3 = 1$)

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

STA 674, RADOE:
Fitting Multiple Linear Regression Models

small difference between ivy and non-ivy, but have to assess significance...in this case it is not significant (P=0.43, t=0.78)

Example: US Graduation Rates

- Separate Intercept Model –

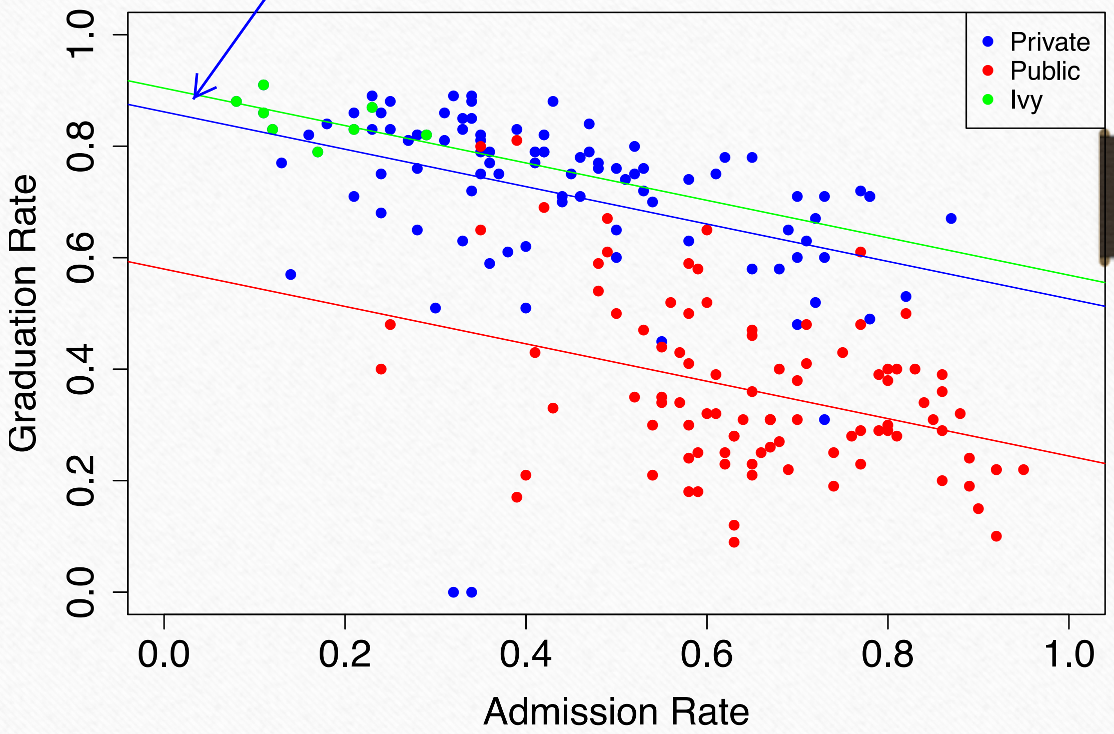
    Private versus Public versus Ivy

Root MSE	0.13990	R-Square	0.6562
Dependent Mean	0.54585	Adj R-Sq	0.6508
Coeff Var	25.62927		

Same as Public vs Private (2 indicators)

Non-ivy vs Ivy

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.57962	0.04230	13.70	<.0001
ADMISRATE	1	-0.33543	0.06077	-5.52	<.0001
privateind	1	0.28201	0.02402	11.74	<.0001
ivyind	1	0.04247	0.05435	0.78	0.4356



STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: US Graduation Rates

error of yhat from actual y

proportion of variability explained by model

adjusted for number of variables; not a statistic but an analytic for comparison

- Summary of models fit:

Model	# of Predictors	*RMSE*	*R*-squared	AdjR^2
Single intercept	1	0.18	0.41	0.40
Public vs. Private	2	0.14	0.66	0.65
Public vs. Private vs. Ivy	3	0.14	0.66	0.65

STA 674, RADOE:
Fitting Multiple Linear Regression Models

- Definition: **Indicator Variables** – An indicator variable is a variable that takes two values, 0 or 1, that distinguish between two qualitatively defined categories.
- **Interpretation**: An indicator models a difference in the ***intercept*** between two groups.
- But what if the slopes look different … ?

STA 674, RADOE:
Fitting Multiple Linear Regression Models

- Definition:

  Interaction: An interaction is formed by the product of two predictors.

- Interpretation

  Interactions model changes in the effect (slope) of one variable on the response as the other variable changes.

  second order response. measure of the change of one variable as another variable is changed

STA 674, RADOE:
Fitting Multiple Linear Regression Models

- Definition:

  Interaction: An interaction is formed by the product of two predictors.

- Interpretation

  Interactions model changes in the effect (slope) of one variable on the response as the other variable changes.

- Suppose that x_1 is a continuous predictor and x_2 is an indicator variable. If:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i,$$

then:

1. The model for Group 1 ($x_2 = 0$) is:

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i$$

2. The model for Group 2 ($x_2 = 1$) is:

$$y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{1i} + e_i$$

intercept difference

slope difference

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: Example: US Graduation Rates

Predictors

- x_1 = admission rate (ADMISRATE)
- x_2 = 0 for public schools, 1 for private schools (PRIVATE)

Complete Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + e_i,$$

Model for Public Schools ($x_{2i} = 0$) is:

$$y_i = \beta_0 + \beta_1 x_{1i} + e_i$$

Model for Private Schools ($x_{2i} = 1$) is:

$$y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{1i} + e_i$$

STA 674, RADOE:
Fitting Multiple Linear Regression Models

Example: Example: US Graduation Rates

Root MSE	0.13977	R-Square	0.6568
Dependent Mean	0.54585	Adj R-Sq	0.6514
Coeff Var	25.60602		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.63554	0.06200	10.25	<.0001
ADMISRATE	1	-0.42084	0.09213	-4.57	<.0001
privateind	1	0.21666	0.07088	3.06	0.0026
ar_x_private	1	0.11558	0.11804	0.98	0.3288

difference in slopes