

Lecture 9

1. Math scores: comparison of subset selection results
2. Math scores: conclusion
3. Continuous and categorical predictors
4. Indicator variables
5. Proc GLM and Proc Reg

1

Model selection “by hand”

Full model:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	71.49920	2.07620	34.44	<.0001
log_lep	1	-0.65463	0.13643	-4.80	<.0001
log_lunch	1	-1.89919	0.27951	-6.79	<.0001
Special_Ed_pct	1	-0.07866	0.03944	-1.99	0.0468
log_mobility	1	-1.53474	0.29114	-5.27	<.0001
drop_out_pct	1	-0.56735	0.15351	-3.70	0.0002
log_8th_grade	1	0.21923	0.30184	0.73	0.4681
log_total_students	1	-0.72168	0.39947	-1.81	0.0716
operating_budget	1	-0.46716	0.30285	-1.54	0.1237
total_budget	1	-0.05989	0.23671	-0.25	0.8004

Within the correlated pairs of predictors related to number of students and budget, drop the predictor with larger p -value.

2

Reduced model 1.

Variable	DF	Parameter	Standard	t Value	Pr > t
		Estimate	Error		
Intercept	1	71.10315	1.99172	35.70	<.0001
log_lep	1	-0.62906	0.13448	-4.68	<.0001
log_lunch	1	-1.96629	0.26792	-7.34	<.0001
Special_Ed_pct	1	-0.06668	0.03809	-1.75	0.0808
log_mobility	1	-1.53766	0.28483	-5.40	<.0001
drop_out_pct	1	-0.58086	0.14566	-3.99	<.0001
log_total_students	1	-0.47905	0.22238	-2.15	0.0318
operating_budget	1	-0.53618	0.15552	-3.45	0.0006

Should we drop anything else?

No: for large observational dataset, keep predictors with $p < .1$ or $p < .2$

3

Automatic subset selection

Proc Reg is one of several regression procedures that offers automatic selection of a smaller model from a full model.

- **Backwards** Starting from full model, sequentially drop predictors with $p >$ specified cutoff.
Done by hand, the most common simple procedure among analysts.
- **Forwards** Start with single predictor with lowest p -value in univariate regression, sequentially add predictors with lowest p -value.
- **Stepwise** Start with Forwards but consider Backwards at each step.
- Maximize a criterion (R^2 , adjusted R^2 , Mallows's C_p):
find models of 1, 2, 3, ..., predictors with largest values of the criterion

4

Reduced Model 2. Backward selection: default value to stay is $p < .1$

```
Proc Reg data=B;
  model mathscore = log_lep log_lunch
    Special_Ed_pct log_mobility drop_out_pct
    log_8th_grade log_total_students operating_budget
    total_budget / selection = backward;
```

5

Backward Elimination: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	71.10464	2.00647	7777.15867	1255.83	<.0001
log_lep	-0.64116	0.13503	139.61877	22.55	<.0001
log_lunch	-1.92395	0.27252	308.66742	49.84	<.0001
Special_Ed_pct	-0.07194	0.03834	21.79910	3.52	0.0613
log_mobility	-1.51180	0.28692	171.93641	27.76	<.0001
drop_out_pct	-0.60028	0.14659	103.84724	16.77	<.0001
log_total_students	-0.48301	0.22450	28.66555	4.63	0.0320
operating_budget	-0.54765	0.15652	75.81050	12.24	0.0005

Bounds on condition number: 1.961, 78.446

All variables left in the model are significant at the 0.1000 level.

So Reduced Model 2 = Reduced Model 1 (7 predictors)

6

Reduced Model 3. Forward selection: default value to include is $p < .5$

No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	log_lunch	1	0.3464	0.3464	213.754	217.81	<.0001
2	log_mobility	2	0.1382	0.4845	84.1232	109.89	<.0001
3	log_lep	3	0.0525	0.5370	36.1297	46.35	<.0001
4	operating_budget	4	0.0177	0.5547	21.2299	16.25	<.0001
5	drop_out_pct	5	0.0132	0.5679	10.6809	12.41	0.0005
6	log_total_students	6	0.0048	0.5727	8.0836	4.59	0.0328
7	Special_Ed_pct	7	0.0037	0.5764	6.5759	3.52	0.0613
8	log_8th_grade	8	0.0005	0.5770	8.0640	0.51	0.4742

8 predictors

```
MODEL . . . / selection=forward slentry= 0.25 ;
```

7

Reduced Model 4. Stepwise selection is second most popular: default value to include or exclude is $p < .15$

Stepwise Selection: Step 7

Variable Special_Ed_pct Entered: R-Square = 0.5764 and C(p) = 6.5759

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	71.10464	2.00647	7777.15867	1255.83	<.0001
log_lep	-0.64116	0.13503	139.61877	22.55	<.0001
log_lunch	-1.92395	0.27252	308.66742	49.84	<.0001
Special_Ed_pct	-0.07194	0.03834	21.79910	3.52	0.0613
log_mobility	-1.51180	0.28692	171.93641	27.76	<.0001
drop_out_pct	-0.60028	0.14659	103.84724	16.77	<.0001
log_total_students	-0.48301	0.22450	28.66555	4.63	0.0320
operating_budget	-0.54765	0.15652	75.81050	12.24	0.0005

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

8

Subset selection using Mallows's C_p

Mallow's C_p measures how well a subset model predicts the observed data (see chapter 8 in Weisberg, *Applied Linear Regression*), and estimates how well the model will predict new observations.

If there are n observations and K predictors in the full model and a subset model has $p \leq K$ predictors, then

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n = (K - p)(F_p - 1) + p.$$

For the full model, $p = K$ so $C_K = K$.

If the omitted variables really have zero coefficients, then the test statistic for this $F_p \approx 1$, so that $C_p \approx p$.

Good prediction models have $C_p \leq p$.

9

With K predictors, there are 2^K subset models. SAS will check all of them and for $K \leq 11$ will print them all.

For our 9 predictors, $2^9 = 512$.

Reduce output by looking at best 5 models at each size:

```
Proc Reg data=B;
  model mathscore = log_lep log_lunch
    Special_Ed_pct log_mobility drop_out_pct
    log_8th_grade log_total_students operating_budget
    total_budget / selection = cp best = 5;
```

C(p) Selection Method

Number in Model	C(p)	R-Square	Variables in Model
7	6.5759	0.5764	log_lep log_lunch Special_Ed_pct log_mobility
8	8.0640	0.5770	log_lep log_lunch Special_Ed_pct log_mobility drop_out_pct log_8th_grade log_total_students operating_budget
6	8.0836	0.5727	log_lep log_lunch log_mobility drop_out_pct log_total_students operating_budget
8	8.5275	0.5765	log_lep log_lunch Special_Ed_pct log_mobility drop_out_pct log_total_students operating_budget total_budget
6	9.1885	0.5716	log_lep log_lunch Special_Ed_pct log_mobility drop_out_pct operating_budget

Good prediction models have $C_p \leq p$.

11

Automatic methods, except for forward selection, gave the same reduced model we chose by hand for this example.

With strongly associated predictors, results are often similar.

Best practice is to choose the subset model(s) by hand with careful consideration of each predictor, and relations among predictors.

Use C_p to compare with your reduced model.

12

Compare fit and regression coefficients of final model with outlier-free data and full data, which has transformations but retains all observations.

```
Proc Reg data=B; ***** reduced model, trimmed data;
  model mathscore = log_lep log_lunch
    Special_Ed_pct log_mobility drop_out_pct
    log_total_students operating_budget;
```

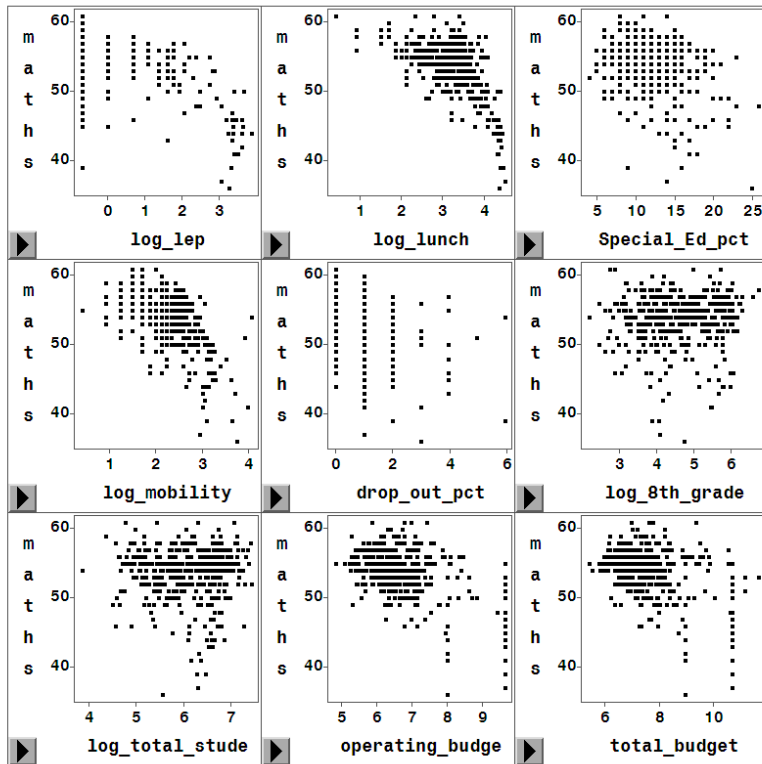
```
Proc Reg data=A; ***** reduced model, full data;
  model mathscore = log_lep log_lunch
    Special_Ed_pct log_mobility drop_out_pct
    log_total_students operating_budget;
```

13

Where are the large differences?

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
<i>trimmed data</i>					
Intercept	1	71.10315	1.99172	35.70	<.0001
log_lep	1	-0.62906	0.13448	-4.68	<.0001
log_lunch	1	-1.96629	0.26792	-7.34	<.0001
Special_Ed_pct	1	-0.06668	0.03809	-1.75	0.0808
log_mobility	1	-1.53766	0.28483	-5.40	<.0001
drop_out_pct	1	-0.58086	0.14566	-3.99	<.0001
log_total_students	1	-0.47905	0.22238	-2.15	0.0318
operating_budget	1	-0.53618	0.15552	-3.45	0.0006
<i>full data</i>					
Intercept	1	60.94595	2.05839	29.61	<.0001
log_lep	1	-0.98326	0.15457	-6.36	<.0001
log_lunch	1	-1.34586	0.27462	-4.90	<.0001
Special_Ed_pct	1	-0.09678	0.04014	-2.41	0.0163
log_mobility	1	0.05344	0.28112	0.19	0.8493
drop_out_pct	1	-1.04380	0.13462	-7.75	<.0001
log_total_students	1	0.48335	0.23904	2.02	0.0438
operating_budget	1	-0.65229	0.17366	-3.76	0.0002

14



15

From here:

Identify small, medium, large districts: what are effects of predictors, controlling district size?

Use model from trimmed data for predictions, LSmeans, etc.

Plots: show full data with results from model.

16

Types of predictors

In regression and ANOVA, we model the response as a function of predictors (explanatory variables, independent variables).

There are two types:

- **Continuous** are real-number values, measured on a continuous scale: height, weight.
- **Categorical** are labels, usually discrete values: gender, country of origin, marital status, high-school graduate, eye color.

Discrete, but often treated as continuous: years of education, number of children, number of automobiles owned

Continuous predictors can be cut into categories.

17

Categorical predictors

In order to compute a regression, categorical predictors must be re-expressed as indicator (0/1) variables: one indicator for each level of the factor.

Gender: M, F

City: A, B, C

gender	male	female	city	city_A	city_B	city_C
M	1	0	B	0	1	0
M	1	0	C	0	0	1
F	0	1	C	0	0	1
F	0	1	A	1	0	0
F	0	1	B	0	1	0
M	1	0	A	1	0	0

18

Design matrix is matrix formed by columns of predictors used in computation:

intercept	male	female	city_A	city_B	city_C	mom_iq
1	1	0	0	1	0	98
1	1	0	0	0	1	115
1	0	1	0	0	1	80
1	0	1	1	0	0	110
1	0	1	0	1	0	84
1	1	0	1	0	0	92

Intercept is a column of 1s

Sum of indicators for a class variable = column of 1s

⇒ Linear dependence

⇒ drop one level from each categorical variable (df = levels - 1)

19

Regress child's IQ on mom's IQ and city of residence (A, B, C):

```
Proc REG data= pubh.child_iq_6470;
  model child_iq = mom_iq city_A city_B city_C;
```

NOTE: Model is not full rank. Least-squares solutions for the parameters are

$$\text{city_C} = \text{Intercept} - \text{city_A} - \text{city_B}$$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	B	27.08798	5.99981	4.51	<.0001
mom_iq	1	0.61071	0.05827	10.48	<.0001
city_A	B	-4.26220	2.21210	-1.93	0.0547
city_B	B	-0.77697	2.03560	-0.38	0.7029
city_C	0	0	.	.	.

20

How do we fit an interaction between city and mom's IQ?

```
data p;
  set pubh.child_iq_6470;
  momiq_cityA = city_A * mom_IQ; calculate interaction terms
  momiq_cityB = city_B * mom_IQ;
  momiq_cityC = city_C * mom_IQ;
```

mom_iq	city	momiq_ cityA	momiq_ cityB	momiq_ cityC
121	C	0	0	121
89	B	0	89	0
115	C	0	0	115
99	C	0	0	99
93	C	0	0	93
108	A	108	0	0

21

```
Proc REG data=p;
  model child_iq = mom_iq city_A city_B city_C
    momiq_cityA momiq_cityB momiq_cityC;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	B	31.05933	9.54857	3.25	0.0012
mom_iq	B	0.57106	0.09431	6.06	<.0001
city_A	B	-6.03885	14.53384	-0.42	0.6780
city_B	B	-11.38060	14.05725	-0.81	0.4186
city_C	0	0	.	.	.
momiq_cityA	B	0.01777	0.14327	0.12	0.9013
momiq_cityB	B	0.10624	0.13929	0.76	0.4460
momiq_cityC	0	0	.	.	.

22

Proc GLM (General Linear Model)

Fits linear models with both continuous and categorical predictors.

Use CLASS statement to identify categorical predictors. Proc GLM makes indicator variables and interactions for you.

```
proc glm data= pubh.child_iq_6470;  
  class city;  
  model child_iq = mom_iq city / solution ;  
  solution requests regression coefficients
```

```
proc glm data= pubh.child_iq_6470;  
  class city;  
  model child_iq = mom_iq city mom_iq*city / solution;
```

23

Model without the interaction term:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
mom_iq	1	36459.75203	36459.75203	109.85	<.0001
city	2	1312.32653	656.16326	1.98	0.1398

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	27.08798161 B	5.99980744	4.51	<.0001
mom_iq	0.61070678	0.05826950	10.48	<.0001
city A	-4.26219697 B	2.21210257	-1.93	0.0547
city B	-0.77697248 B	2.03559961	-0.38	0.7029
city C	0.00000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse

24

Model including interaction term:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
mom_iq	1	36216.74010	36216.74010	108.77	<.0001
city	2	219.46259	109.73129	0.33	0.7194
mom_iq*city	2	213.09066	106.54533	0.32	0.7263

Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		31.05933001 B	9.54856656	3.25	0.0012
mom_iq		0.57106014 B	0.09430752	6.06	<.0001
city	A	-6.03885405 B	14.53383718	-0.42	0.6780
city	B	-11.38059525 B	14.05724822	-0.81	0.4186
city	C	0.00000000 B	.	.	.
mom_iq*city	A	0.01777199 B	0.14326676	0.12	0.9013
mom_iq*city	B	0.10624020 B	0.13928502	0.76	0.4460
mom_iq*city	C	0.00000000 B	.	.	.

25

Proc REG and Proc GLM

In regression, we're interested in slopes.

In ANOVA, we're interested in means.

These are really the same method (OLS).

General linear model has both categorical (treatment or study group) and continuous predictors.

Usual aim: compare means in some categories, adjusted for continuous predictors.

26

Proc Reg has no CLASS statement, so indicators and interactions must be calculated.

Advantages: automatic stepwise selection of variables, diagnostics for collinearity

Proc GLM has a CLASS statement, and options for computing and comparing adjusted means

Using ODS, both procedures can draw many graphics

Both will output a dataset containing predicted values + diagnostics