Lecture 11

- Interaction models and qualitative predictor
- Redundant variables
- Model selection
- Partial regression plots

Interaction Models

- With several explanatory variables, we need to consider the possibility that the effect of one variable depends on the value of another variable
- Special cases
  - One binary variable and one continuous variable
  - Two continuous variables

One binary variable and one continuous variable

- X_1 has values 0 and 1 corresponding to two different groups
- X_2 is a continuous variable
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \xi$
- For $X_1 = 0$, $Y = \beta_0 + \beta_2 X_2 + \xi$
- For $X_1 = 1$, $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \xi$

One binary and one continuous (2)

- For $X_1 = 0$, $Y = \beta_0 + \beta_2 X_2 + \xi$
- For $X_1 = 1$, $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \xi$
- H_0: $\beta_1 = \beta_3 = 0$ tests the hypothesis that the lines are the same
- H_0: $\beta_1 = 0$ tests equal intercepts
- H_0: $\beta_3 = 0$ tests equal slopes

Example

- Y is number of months for an insurance company to adopt an innovation
- X_1 is the size of the firm (a continuous variable)
- X_2 is the type of firm (a qualitative or categorical variable)

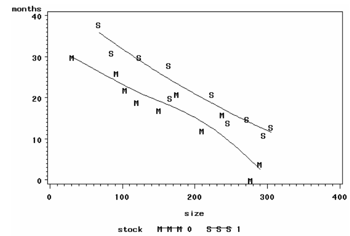
A question

- X_2 (the type of firm) has the value 0 for a mutual fund and 1 for a stock fund
- We ask whether or not stock firms adopt the innovation slower or faster than mutual firms
- We ask the question across all firms, regardless of size

Plot the data

```
symbol1 v=M i=sm70 c=green;
symbol2 v=S i=sm70 c=blue;
proc sort data=a1;
   by stock size;
proc gplot data=a1;
   plot months*size=stock;
run;
```

Two symbols



Interaction effects

- Interaction expresses the idea that the effect of one explanatory variable on the response depends on another explanatory variable
- In our example, this would mean that the slope of the line depends on the type of firm

Are both lines the same ?

- Are intercepts and slopes the same ? (test statement)

```
Data a1; set a1;
sizestoc=size*stock;
proc reg data=a1;
   model months=
     size stock sizestoc;
     test stock, sizestoc;
run;
```

Output (Overall ANOVA)

F Value	Pr > F
45.49	<.0001

R-Square 0.8951

Output (test statement) Are both lines the same ?

Test 1 Results for Dependent Variable months

Source	*DF*	*MS*	*F*	*P>F*
Num	*2*	*158.13*	*14.34*	*0.0003*
Den	*16*	*11.02*		

Output (3)
How are they different ?

Variable	t Value	Pr > |t|
Intercept	13.86	<.0001
size	-7.78	<.0001
stock	2.23	0.0408
sizestoc	-0.02	0.9821

Two parallel lines

```
proc reg data=a1;
   model months=size stock;
run;
```

Output

Source	DF	F	Pr > F
Model	2	72.50	<.0001
Error	17		
Total	19		

Output (2)

Root MSE	3.22113
R-Square	0.8951
Dependent Mean	19.40000
Adj R-Sq	0.8827
Coeff Var	16.60377

Output (3)

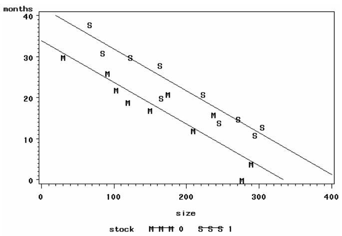
Var	DF	Par Est	St Err	t	P
Int	1	33.87	1.8	18.68	<.0001
size	1	-0.10	0.0	-11.44	<.0001
stock	1	8.05	1.4	5.52	<.0001

Int for stock firms is
33.87+8.05 = 41.92

Plot the two lines

```
symbol1 v=M i=rl c=green;
symbol2 v=S i=rl c=blue;
proc gplot data=a1;
 plot months*size=stock;
run;
```

The plot



Two continuous variables

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \xi$
- $Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \xi$
- $Y = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \xi$

Constrained regression

- We may want to put a linear constraint on the regression coefficients, e.g. $\beta_1 = 1$, or $\beta_1 = \beta_2$
- We can do this by redefining our explanatory variables (data step)
- Or we can use the RESTRICT statement in proc reg (e.g. restrict size=0; or restrict size=5*stock;)

Redundant variables

- **data** a1;
- infile 'u:/.www/STAT512/data/example1.txt'; input x1 x2 x3;
- **proc corr** data=a1;
- var x1 x2 x3;
- **run**;
- **data** a2; set a1; y1=x1+normal(**0**);
- **run**;
- **proc reg** data=a2;
- model1: model y1=x1;
- model2: model y1=x1 x2;
- model3: model y1=x1 x2 x3;
- **run**;

---

x1	x2	x3
4	2	-1
4	2	1
4	3	-1
4	3	1
6	2	-1
6	2	1
6	3	-1
6	3	1

---

	x1	x2	x3
x1	1.00000	0.00000	0.00000
		1.0000	1.0000
x2	0.00000	1.00000	0.00000
	1.0000		1.0000
x3	0.00000	0.00000	1.00000
	1.0000	1.0000	

- Var slope std t p-value
- x1 1.28612 0.50195 2.56 0.0428

- x1 1.28612 0.52685 2.44 0.0586
- x2 -0.70382 1.05371 -0.67 0.5338

- x1 1.28612 0.58875 2.18 0.0943
- x2 -0.70382 1.17751 -0.60 0.5822
- x3 -0.03677 0.58875 -0.06 0.9532

Conclusion

- **Redundant variables increase the error and decrease the power of detection of important coefficients.**

Variable Selection

- **We want to choose a model that includes a subset of the available explanatory variables**
- **Two separate problems**
  - **How many explanatory variables should we use (subset size)**
  - **Given the subset size, which variables should we choose**

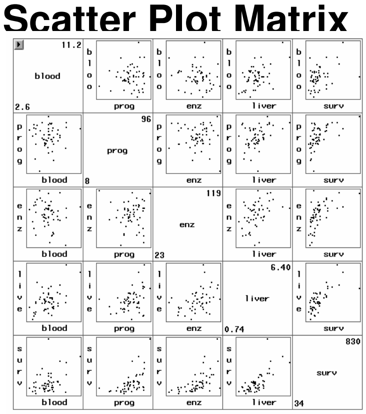
Example

- **Y is survival time**
- **X's are**
  - **Blood clotting score**
  - **Prognostic index**
  - **Enzyme function test**
  - **Liver function test**

Example (2)

- **n = 54 patients**
- **Diagnostics suggest that Y should be transformed with a log**
- **Start with the usual plots and descriptive statistics**

Scatter Plot Matrix



The two problems in variable selection

- To determine an appropriate subset size you may use e.g. C_p, SBC or AIC
- For comparing models with the same number of variables, we use R^2

C_p

- The basic idea is to compare subset models with the full model
- A subset model is good if there is not substantial bias in the predicted values (relative to the full model)
- Mean squared error - $E(\hat{Y}_i - \mu_i)^2 = B_i$
- C_p is an estimator of $\sum_{i=1}^{n} B_i / \sigma^2$

C_p

$$C_p = \frac{SSE_p}{MSE(F)} - (n - 2p)$$

Use of C_p

- p is the number of regression coefficients including the intercept (this is consistent with the notation we have been using)
- A model is good according to this criterion if C_p is close to or smaller then p
- Pick the smallest model for which
- C_p is close to p or the one for which C_p is the smallest

SBC and AIC

Chose the model for which log(likelihood) - penalty for the dimension is maximal

AIC – minimize $n\log\left(\frac{SSE_p}{n}\right) + 2p$

- SBC – minimize $n\log\left(\frac{SSE_p}{n}\right) + p\log(n)$

Ordering models of the same subset size

- use R^2
- This approach can lead us to consider several models (subsets) that give us approximately the same predicted values
- We may need to apply knowledge of the subject matter to make a final selection

Proc reg

```
proc reg data=a1;
   model lsurv=
     blood prog enz liver/
     selection=rsquare cp aic
     sbc b best=3;
run;
```

Number in Model	R-Square	C(p)	AIC	SBC
1	0.4276	66.4889	-103.8269	-99.84889
1	0.4215	67.7148	-103.2615	-99.28357
1	0.2208	108.5558	-87.1781	-83.20011
2	0.6633	20.5197	-130.4833	-124.51634
2	0.5995	33.5041	-121.1126	-115.14561
2	0.5486	43.8517	-114.6583	-108.69138
3	0.7573	3.3905	-146.1609	-138.20494
3	0.7178	11.4237	-138.0232	-130.06723
3	0.6121	32.9320	-120.8442	-112.88823
4	0.7592	5.0000	-144.5895	-134.64461

Model	R-Square	Intercept	blood	prog	enz	liver
1	0.4276	5.26426	.	.	0.01512	.
1	0.4215	5.61218	.	.	.	0.29819
1	0.2208	5.56613	.	0.01367	.	.
2	0.6633	4.35058	.	0.01412	0.01539	.
2	0.5995	5.02818	.	.	0.01073	0.20945
2	0.5486	4.54623	0.10792	.	0.01634	.
3	0.7573	3.76618	0.09546	0.01334	0.01645	.
3	0.7178	4.40582	.	0.01101	0.01261	0.12977
3	0.6121	4.78168	0.04482	.	0.01220	0.16360
4	0.7592	3.85195	0.08368	0.01266	0.01563	0.03216

- **data** a1;
- infile 'u:/.www/STAT512/data/ch07ta01.txt';
- input x1 x2 x3 y;
- **run**;
- **proc reg** data=a1;
- model y=
- x1 x2 x3/
- selection=rsquare cp aic
- sbc b;
- **run**;

Model	R-Square	C(p)	AIC	SBC
1	0.7710	2.4420	38.7080	40.69942
1	0.7111	7.2703	43.3590	45.35045
1	0.0203	62.9128	67.7823	69.77373
2	0.7862	3.2242	39.3417	42.32891
2	0.7781	3.8773	40.0860	43.07321
2	0.7757	4.0657	40.2957	43.28293
3	0.8014	4.0000	39.8672	43.85009

Model	R-Square	Intercept	x1	x2	x3
1	0.7710	-23.63449	.	0.85655	.
1	0.7111	-1.49610	0.85719	.	.
1	0.0203	14.68678	.	.	0.19943
2	0.7862	6.79163	1.00058	.	-0.43144
2	0.7781	-19.17425	0.22235	0.65942	.
2	0.7757	-25.99695	.	0.85088	0.09603
3	0.8014	117.08469	4.33409	-2.85685	-2.18606

Variable Selection

- **Additional proc reg model statement options useful in variable selection**
  - **INCLUDE=n forces the first *n* explanatory variables into all models**
  - **BEST=n limits the output to the best *n* models of each subset size**
  - **MAXSTEP=n limits the number of steps in forward, backward and stepwise methods.**

- **START=n for Cp option limits output to models that include at least *n* explanatory variables,**
- **For stepwise it begins the search process with first n explanatory variables specified in the model statement**

Other approaches

- **Maximize adjusted R^2 (ADJRSQ)**
- **PRESS (prediction SS)**
  - **For each case i**
  - **Delete the case and predict Y using a model based on the other n-1 cases**
  - **Look at the SS for observed minus predicted**

Other approaches (2)

- **Step type procedures**
  - **Forward selection (Step up)**
  - **Backward elimination (Step down)**
  - **Stepwise (forward selection with a backward glance)**

Backward elimination

- **data** a1;
- infile 'u:/.www/STAT512/data/ch07ta01.txt';
- input x1 x2 x3 y;
- **run**;
- **proc reg** data=a1;
- model y=x1 x2 x3/selection=b;
- **run**;

- Backward Elimination: Step 0
- All Variables Entered: R-Square = 0.8014 and C(p) = 4.0000

Var	coef	std err	t	p-value
Intercept	117.08469	99.782	1.38	0.2578
x1	4.33409	3.015	2.07	0.1699
x2	-2.85685	2.582	1.22	0.2849
x3	-2.18606	1.595	1.88	0.1896

- Backward Elimination: Step 1
- Variable x2 Removed: R-Square = 0.7862 and C(p) = 3.2242
- Intercept 6.79163 4.48829 2.29 0.1486
- x1 1.00058 0.12823 60.89 <.0001
- x3 -0.43144 0.17662 5.97 0.0258
- All variables left in the model are significant at the 0.1000 level.
- Summary of Backward Elimination
- Var rem. R^2 C(p) F p-value
- x2 0.7862 3.2242 1.22 0.2849

Forward selection

- **proc reg** data=a1;
- model y=x1 x2 x3/selection=f;
- **run**;

- Forward Selection: Step 1
- Variable x2 Entered: R-Square = 0.7710 and C(p) = 2.4420
- Var coef std t p
- Intercept -23.63449 5.65741 17.45 0.0006
- x2 0.85655 0.11002 60.62 <.0001

- Forward Selection: Step 2
- Variable x1 Entered: R-Square = 0.7781 and C(p) = 3.8773
- Intercept -19.17425 8.36064 5.26 0.0348
- x1 0.22235 0.30344 0.54 0.4737
- x2 0.65942 0.29119 5.13 0.0369

- Forward Selection: Step 3
- Variable x3 Entered: R-Square = 0.8014 and C(p) = 4.0000
- Var coef std err t p
- Intercept 117.08469 99.78240 1.38 0.2578
- x1 4.33409 3.01551 2.07 0.1699
- x2 -2.85685 2.58202 1.22 0.2849
- x3 -2.18606 1.59550 1.88 0.1896

- All variables have been entered into the model.
- Summary of Forward Selection

Step var R^2 c(p) F p
- 1 x2 0.7710 2.4420 60.62 <.0001
- 2 x1 0.7781 3.8773 0.54 0.4737
- 3 x3 0.8014 4.0000 1.88 0.1896

Stepwise selection

- **proc reg** data=a1;
- model y=x1 x2 x3/selection=stepwise;
- **run**;
- quit;

- Stepwise Selection: Step 1
- Variable x2 Entered: R-Square = 0.7710 and C(p) = 2.4420

Var	coef	std err	t	p
Intercept	-23.63449	5.65741	17.45	0.0006
x2	0.85655	0.11002	60.62	<.0001

- 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.

- Summary of Stepwise Selection

step	var ent	R^2	C(p)	F	p
1	x2	0.7710	2.4420	60.62	<.0001

SAS Defaults

- **SLstay (significance level to remove a variable from a model) = 0.1 for backward elimination, 0.15 for stepwise selection**
- **SLenter(significance level to add a new variable into a model) = 0.5 for forward selection, 0.15 for stepwise selection**

Partial regression plots

- **Also called added variable plots or adjusted variable plots**
- **One plot for each X_i**

Partial regression plots (2)

- **Consider X_1**
  - **Use the other X's to predict Y**
  - **Use the other X's to predict X_1**
  - **Plot the residuals from the first regression vs the residuals from the second regression**

Partial regression plots (3)

- **These plots show the strength of relatioship between Y and X_i in the full model. They can also detect**
  - **Nonlinear relationships**
  - **Heterogeneous variances**
  - **Outliers**

Example

- Y is amount of life insurance
- X_1 is average annual income
- X_2 is a risk aversion score
- n = 18 managers

Create a data set

```
data a1;
infile `h:/STAT512/ch10ta01.txt';
 input income risk insur;
```

The partial option with proc reg

```
proc reg data=a1;
   model insur=income risk
          /partial;
run;
```

Output

```
Source    DF   F Value   Pr > F

Model      2    542.33   <.0001
Error     15
C Total   17
```

Output (2)

```
Root MSE    12.66267
R-Square      0.9864
```

Output (3)

```
       Par  St
Var    Est Err     t Pr > |t|

Int   -205  11 -18.06 <.0001
income 6.2 .20  30.80 <.0001
risk   4.7 1.3   3.44 0.0037
```

Output 4

- The partial option on the model statement in proc reg generates graphs in the output window
- These are ok for some purposes but we prefer to use proc gplot with a smooth
- To generate these plots we do it the hard way

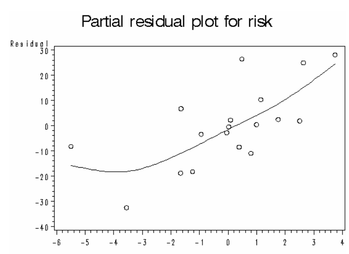
Partial regression plots – the hard way

```
Title1 'Partial residual
     plot for risk';
proc reg data=a1;
  model insur risk = income;
  output out=a2 r=resins resris;
```

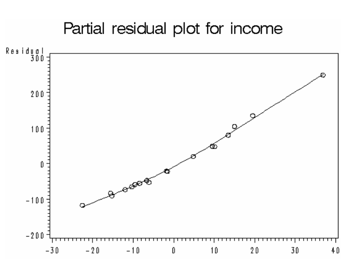
Partial regression plots – the hard way (2)

```
symbol1 v=circle i=sm70s;
proc gplot data=a2;
   plot resins*resris;
run;
```

The plot for risk



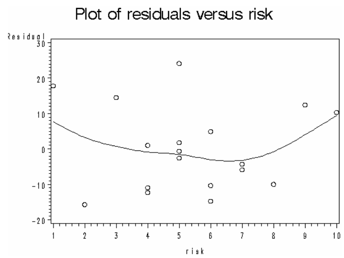
Similar code for income



Plot the residuals vs risk

```
proc reg data=a1;
  model insur= risk income;
   output out=a2 r=resins;
symbol1 v=circle i=sm70;
Title1 'Plot of residuals
  versus risk';
proc sort data=a2; by risk;
proc gplot data=a2;
   plot resins*risk;
run;
```

The graph



Plot the residuals vs income

```
Title1 'Plot of residuals
  versus income';
proc sort data=a2; by income;
proc gplot data=a2;
   plot resins*income;
run;
```

Plot the residuals vs income