STOR 455
STATISTICAL METHODS I

Jan Hannig

Exam 2

- November 2, in class
- Multiple choice - 28 questions
  - Bring your own bubble sheet, pencil and calculator
  - Closed book, closed notes. No computer!
  - You can bring one regular sheet of paper with formulas.
  - Tables will be provided
- Post your questions on blackboard!

Mileage example

- Response variable: Miles per gallon
- Explanatory variable: Speed in miles per hour
- Is the relationship linear?
- Can we use MLR to model it?

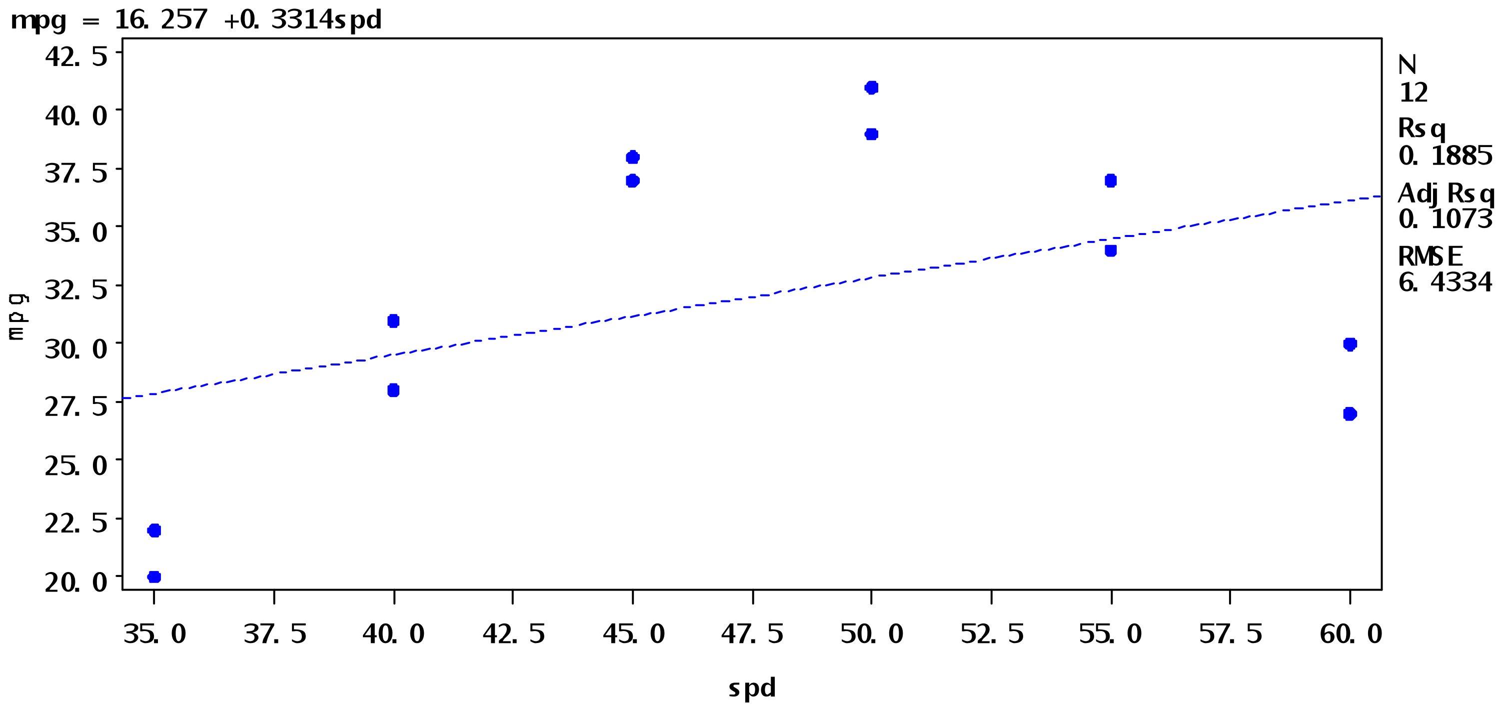
Do it in SAS

```
data mileage;
   infile 'C:/…/mileage.txt';
   input mpg spd;
proc print data=mileage;
run;

symbol1 v=dot h=.8 c=blue;
proc reg data=mileage;
   model mpg = spd;
   plot mpg * spd;
   run;
```

Obs	mpg	spd
1	22	35
2	20	35
3	28	40
4	31	40
5	37	45
6	38	45
7	41	50
8	39	50
9	34	55
10	37	55
11	27	60
12	30	60

Do it in SAS


mpg = 16. 257 +0. 3314spd
42. 5
40. 0
37. 5
35. 0
32. 5
30. 0
27. 5
25. 0
22. 5
20. 0
mpg
35. 0
37. 5
40. 0
42. 5
45. 0
47. 5
50. 0
52. 5
55. 0
57. 5
60. 0
spd
N
12
Rsq
0. 1885
Adj Rsq
0. 1073
RMSE
6. 4334

Polynomial regression

- Useful when response function nonlinear
- Add quadratic, cubic or higher order terms in the model by defining squares, cubes, etc. in a data step and using these as predictors in a multiple regression

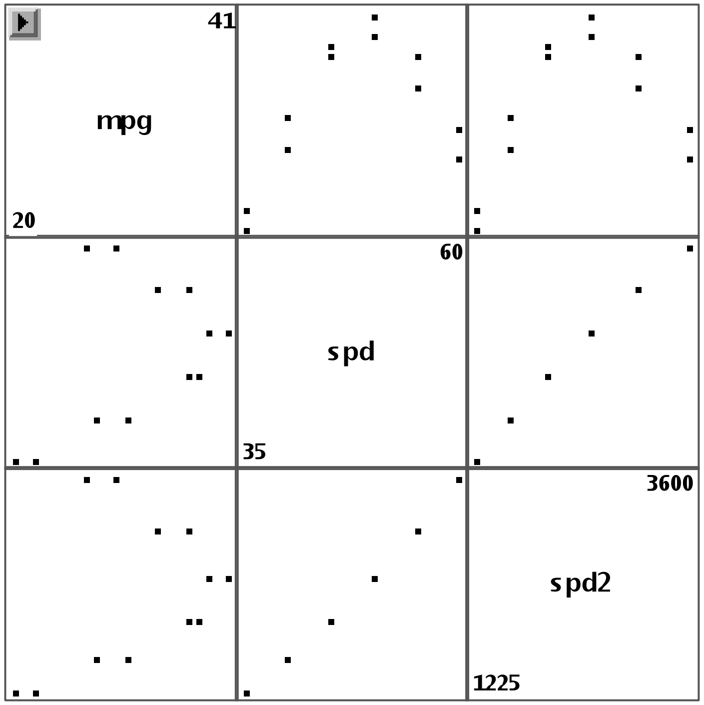
Do it in SAS

```
* creat quadratic term;
data mileage;
   set mileage;
spd2=spd*spd;

%include "C:\...
   \scatter.sas";
%scatter(data = mileage, var
   = mpg spd spd2);

proc corr data=mileage;
run;

proc reg data=mileage;
   model mpg=spd spd2;
   run;
```



Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	483.16786	241.58393	81.03	<.0001
Error	9	26.83214	2.98135		
Corrected Total	11	510.00000			

Root MSE	1.72666	R-Square	0.9474
Dependent Mean	32.00000	Adj R-Sq	0.9357
Coeff Var	5.39581		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-182.58214	17.67703	-10.33	<.0001
spd	1	8.98321	0.76156	11.80	<.0001
spd2	1	-0.09107	0.00799	-11.39	<.0001

Polynomial regression II

- Multicollinearity problem
- Remedy: square centered value x of X (SAS proc standard)
- Hierarchical approach to fitting
- Derive s.d. for regression coefficient of X
- Can do this with more than one explanatory variable

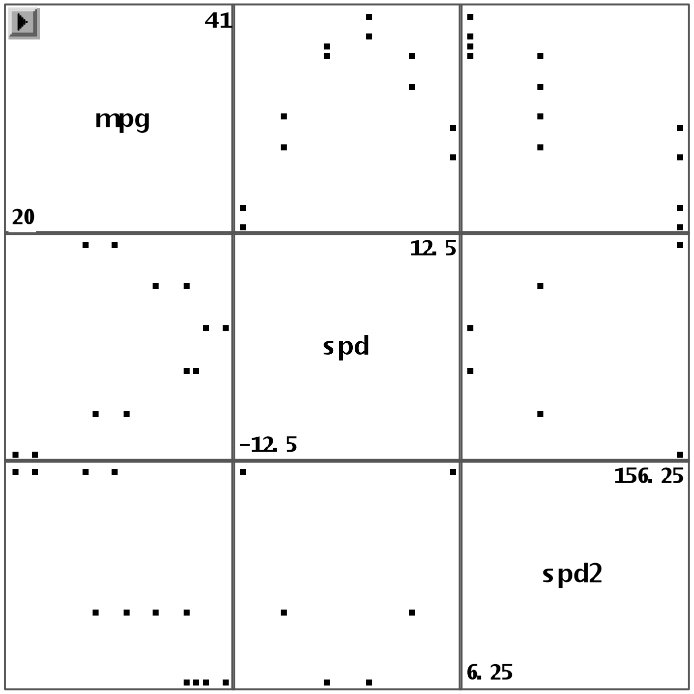
Do it in SAS

```
* Centered value;
proc standard data=mileage
   out=m2 mean=0;
   var spd;

data m2; set m2;
   spd2=spd*spd;

%include "D:\Stat101\SAS
   Macro\scatter.sas";
%scatter(data = m2, var =
   mpg spd spd2);

proc reg data=m2;
   model mpg=spd spd2 /covb;
   run;
```



Do it in SAS

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	38.64063	0.76689	50.39	<.0001
spd	1	0.33143	0.05837	5.68	0.0003
spd2	1	-0.09107	0.00799	-11.39	<.0001

Covariance of Estimates

Variable	Intercept	spd	spd2
Intercept	0.5881177145	0	-0.004658358
spd	0	0.0034072562	0
spd2	-0.004658358	0	0.0000638861

Do it in SAS

```
*Creat cubic term;
data m2; set m2;
      spd3=spd2*spd;

*Test cubic term;
proc reg data=m2;
      model mpg=spd spd2 spd3;
      test spd3;
      run;
```

Test 1 Results for Dependent Variable mpg

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	2.33611	0.76	0.4079
Denominator	8	3.06200		

Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	485.50397	161.83466	52.85	<.0001
Error	8	24.49603	3.06200		
Corrected Total	11	510.00000			

Root MSE	1.74986	R-Square	0.9520
Dependent Mean	32.00000	Adj R-Sq	0.9340
Coeff Var	5.46831		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	38.64063	0.77719	49.72	<.0001
spd	1	0.46703	0.16614	2.81	0.0228
spd2	1	-0.09107	0.00810	-11.24	<.0001
spd3	1	-0.00107	0.00123	-0.87	0.4079

Indicator Variables and Qualitative Variables

- X_i = 1 or 0 to indicate which of the two classes the ith obs. belongs to (i.e., male or female, treatment or control, etc.).
- Also called dummy variables or binary variables.
- Qualitative variable with c classes can be represented by c-1 indicator variables. Example: Education level (HS, College, MS, PHD)

Interaction Models

- If model includes more than one explanatory variables, need to consider possible interaction
- Interaction: the effect of one variable depends on the value of another variable
- Reinforcement or interference
- Implementation: create cross-product in data step

Example of interaction

- Predict yield using fertilizer and raining days (two continuous)
- Predict salary of computer professionals using education, experience and management responsibility (one binary one continuous)

Two continuous variables

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

One binary 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 $X_1 = 1$, $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \xi$

Power cell Example

- Response variable is the life (in cycles) of a power cell
- Explanatory variables are
  - Charge rate (3 levels)
  - Temperature (3 levels)
- This is a designed experiment
- We will use a model with polynomials and test interactions.

Do it in SAS

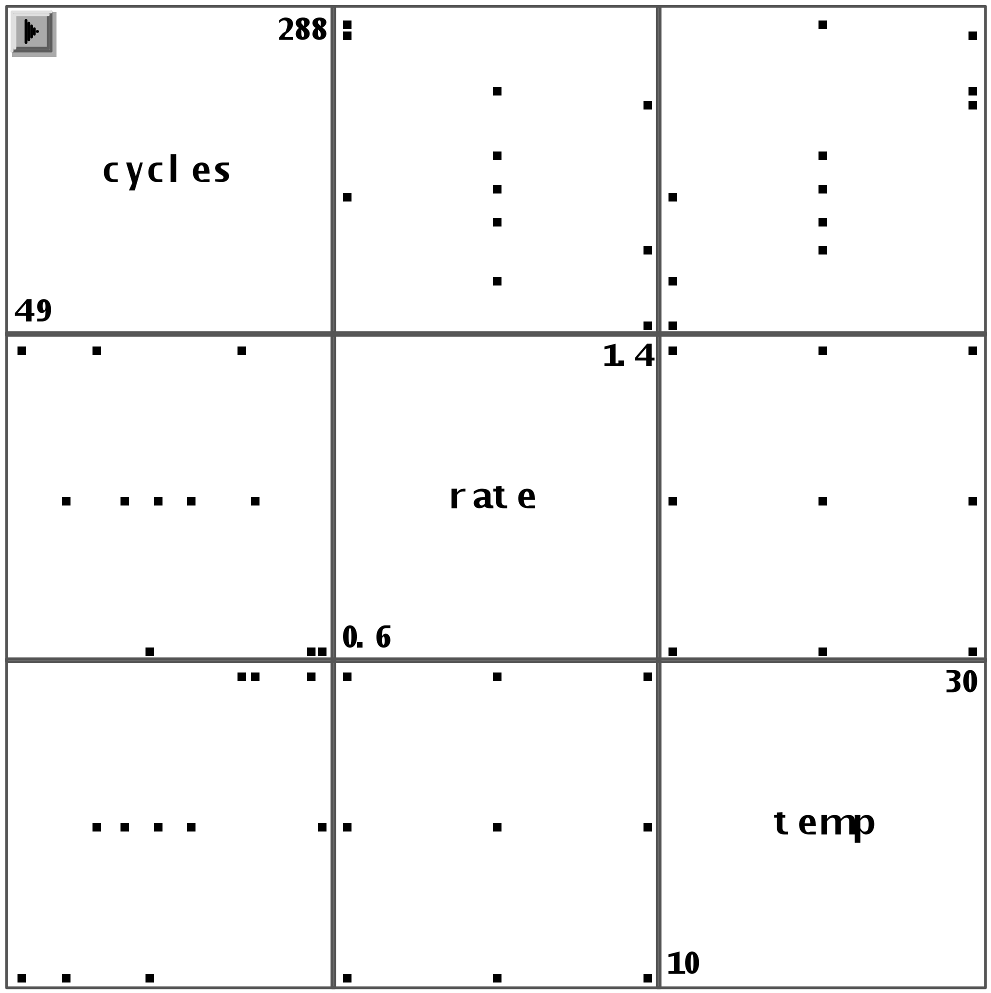
```
Data cell;
   infile 'C:\…
  \powercell.txt';
   input cycles rate temp;
run;

* Making scatter plot using
  macro;

%include "C:\…
  \scatter.sas";
%scatter(data = cell, var =
  cycles rate temp);
```

cycles	rate	temp
150	0.6	10
86	1.0	10
49	1.4	10
288	0.6	20
157	1.0	20
131	1.0	20
184	1.0	20
109	1.4	20
279	0.6	30
235	1.0	30
224	1.4	30

Do it in SAS



Do it in SAS

```
*create second order terms;
Data cell; set cell;
  rate2=rate*rate;
  temp2=temp*temp;
  rt=rate*temp;
*fit model with interaction and
 quadratic term;
Proc reg data=cell;
   model cycles=rate temp rate2 temp2
 rt;
run;
```

Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	55366	11073	10.57	0.0109
Error	5	5240.43860	1048.08772		
Corrected Total	10	60606			

Root MSE	32.37418	R-Square	0.9135
Dependent Mean	172.00000	Adj R-Sq	0.8271
Coeff Var	18.82220		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	337.72149	149.96163	2.25	0.0741
rate	1	-539.51754	268.86033	-2.01	0.1011
temp	1	8.91711	9.18249	0.97	0.3761
rate2	1	171.21711	127.12550	1.35	0.2359
temp2	1	-0.10605	0.20340	-0.52	0.6244
rt	1	2.87500	4.04677	0.71	0.5092

Do it in SAS

```
* Standardize rate
  and temp;
Data c2;
   set cell;
   srate=rate;
  stemp=temp;
   keep cycles srate
  stemp;

Proc standard data=c2
   out=c2 mean=0
  std=1;
   var srate stemp;
```

```
Data c2; set c2;
  srate2=srate*srate;
  stemp2=stemp*stemp;
  srt=srate*stemp;
*test quadratic terms
  and interaction;
Proc reg data=c2;
  model cycles=srate
  stemp srate2 stemp2
  srt;
  test srate2,
  stemp2, srt;
run;
```

Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	55366	11073	10.57	0.0109
Error	5	5240.43860	1048.08772		
Corrected Total	10	60606			

Root MSE	32.37418	R-Square	0.9135
Dependent Mean	172.00000	Adj R-Sq	0.8271
Coeff Var	18.82220		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	162.84211	16.60761	9.81	0.0002
srate	1	-43.24831	10.23762	-4.22	0.0083
stemp	1	58.48205	10.23762	5.71	0.0023
srate2	1	16.43684	12.20405	1.35	0.2359
stemp2	1	-6.36316	12.20405	-0.52	0.6244
srt	1	6.90000	9.71225	0.71	0.5092

Test 1 Results for Dependent Variable cycles

Source	DF	Mean Square	F Value	Pr > F
Numerator	3	819.96491	0.78	0.5527
Denominator	5	1048.08772		

Insurance Company Example

- Innovation in the insurance industry adopted at different speed by different firms.
- Y: number of months for an insurance company to adopt an innovation.
- X_1: the size of the firm in terms of total assets (a continuous variable).
- X_2 : the type of the firm: stock or mutual (a qualitative or categorical variable)

Insurance Company Example

- X_2 (the type of firm) equals 0 for a mutual fund and 1 for a stock fund.
- Q1: Do larger companies adopt innovation faster or slower?
- Q2: Do stock firms adopt the innovation slower or faster than mutual firms?
- Does answer to Q1 depend on the type or the firm? Does answer to Q2 depend on the size?

Import the Data

```
data insu;
   infile 'C:\...
  \Ch08ta02.txt';
   input y x1 x2;
   label  y = 'Months'
          x1 = 'Size'
         x2 = 'Firm
  Indicator';

proc print data=insu;
run;

proc reg data = insu;
  model y = x1 x2/ clb;
run;
```

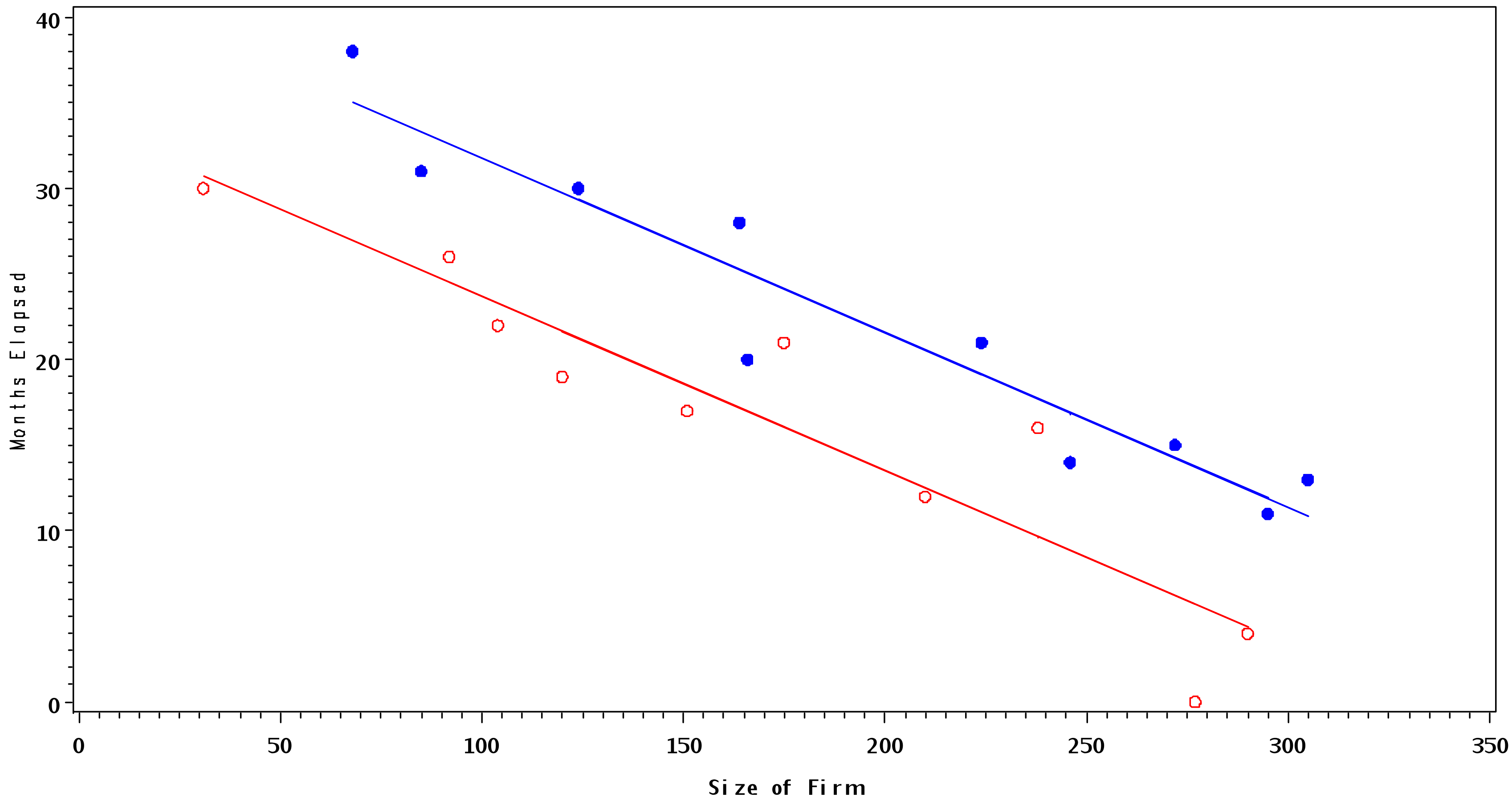
Obs	Months	size	type
1	17	151	0
2	26	92	0
3	21	175	0
4	30	31	0
5	22	104	0
6	0	277	0
7	12	210	0
8	19	120	0
9	4	290	0
10	16	238	0
11	28	164	1
12	15	272	1
13	11	295	1
14	38	68	1
15	31	85	1
16	21	224	1
17	20	166	1
18	13	305	1
19	3	124	1
20	4	246	1

Plot the Data

```
* plot the data;
proc reg data = insu1
  noprint;
  model y = z1 ;
  output out = temp1 p =
  p1;
run;
proc reg data = temp1
  noprint;
  model y = z2;
  output out=temp p= p2;
run;
```

```
symbol1 c=red v=circle;
symbol2 c=blue v=dot
  i=none;
symbol3 i=join v=none
  c=red;
symbol4 i=join v=none
  c=blue;
axis1 order=(0 to 350 by
  50)label=('Size of
  Firm');
axis2 label=(angle = 90
  'Months Elapsed');
proc gplot data = temp;
  plot y1*z1  y2*z2 p1*z1
  p2*z2 / overlay haxis =
  axis1 vaxis=axis2;
run;
```

Do it in SAS



Months Elapsed
40
30
20
10
0
0
50
100
150
200
250
300
350
Size of Firm

Do it in SAS

```
* fit the model;
data insu;
  set insu;
  x1x2 = x1*x2;
run
proc reg data = insu;
  model y = x1 x2 /clb;
run;
;
```

Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1504.41333	752.20667	72.50	<.0001
Error	17	176.38667	10.37569		
Corrected Total	19	1680.80000			

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

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	33.87407	1.81386	18.68	<.0001
x1	Size	1	-0.10174	0.00889	-11.44	<.0001
x2	Firm Indicator	1	8.05547	1.45911	5.52	<.0001

Parameter Estimates

Variable	Label	DF	95% Confidence Limits	
Intercept	Intercept	1	30.04716	37.70098
x1	Size	1	-0.12050	-0.08298
x2	Firm Indicator	1	4.97703	11.13391

Interpretation of Coefficients

- $Y = 33.87 - 0.10X_1 + 8.06\, X_2$
- For both stock firms and mutual firms, larger firms adopt innovation faster. One more million dollar in assets corresponds to 0.1 month faster in adopting innovations.
- For firms of similar size, stock firms adopt innovations about 8 months later than mutual firms.

Check Interaction

- If the linear relationship between Y and X_1 depends on the type X_2, we say there are interaction between X_1 and X_2.
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \xi$
- For $X_1 = 0$, $Y = \beta_0 + \beta_2 X_2 + \xi$
- For $X_1 = 1$, $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \xi$

Do it in SAS

```
* fit the model and test interaction;
data insu;
  set insu;
  x1x2 = x1*x2;
run;
proc reg data = insu;
  model y = x1 x2 x1x2;
  test x1x2;
run;
```

Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1504.41904	501.47301	45.49	<.0001
Error	16	176.38096	11.02381		
Corrected Total	19	1680.80000			

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	33.83837	2.44065	13.86	<.0001
x1	Size	1	-0.10153	0.01305	-7.78	<.0001
x2	Firm Indicator	1	8.13125	3.65405	2.23	0.0408
x1x2		1	-0.00041714	0.01833	-0.02	0.9821

Test 1 Results for Dependent Variable y

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.00571	0.00	0.9821
Denominator	16	11.02381		

Constrained regression

- We may want to put a linear constraint on the regression coefficients, e.g. $\beta_1 = 1$, or $\beta_1 = \beta_2$
- Method I: redefine explanatory variables in data step
- Method II: use the RESTRICT statement in proc reg