Lecture 7

- Extra Sums of Squares with applications
- Partial correlations
- Standardized regression coefficients

General Linear Tests

- A different way to look at the comparison of models
- Look at the difference
  - in SSE
  - In SSM
- Because SSM+SSE=SST, these two ways are equivalent

General Linear Tests (2)

- Models we compare are hierarchical in the sense that one includes all of the explanatory variables of the other
- We can compare models with different explanatory variables
  - X_1, X_2 vs X_1
  - X_1, X_2, X_3, X_4, X_5 vs X_1, X_2, X_3
- Note first includes all Xs of second

General Linear Tests (3)

- We will get an F test that compares the two models
- We are testing a null hypothesis that the regression coefficients for the *extra* variables are all zero
- For X_1, X_2, X_3, X_4, X_5 vs X_1 , X_2 , X_3
  - H_0: $\beta_4 = \beta_5 = 0$
  - H_1: β_4 and β_5 are not both 0

General Linear Tests (4)

- F=((SSE(R) - SSE(F))/(dfE(R) - dfE(F)))/ MSE(F)
- Degrees of freedom for the F statistic are the number of *extra* variables and the dfE for the model with larger number of explanatory variables
- Suppose n=100 and we compare models with X_1, X_2, X_3, X_4, X_5 vs X_1 , X_2 , X_3
- Numerator df is 2
- Denominator df is n-6 = 94

Notation for Extra SS

- SSE(X_1, X_2, X_3, X_4, X_5) is the SSE for the *full* model
- SSE(X_1, X_2, X_3) is the SSE for the *reduced* model
- SSE(X_4, X_5 | X_1, X_2, X_3) is the difference
- SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4, X_5), or
- SSM(X_1, X_2, X_3, X_4, X_5) - SSM(X_1, X_2, X_3)

F test

- Numerator is (SSE(X_4, X_5 | X_1, X_2, X_3))/2
- Denominator is MSE(X_1, X_2, X_3, X_4, X_5)
- F ~ F(2, n-6)
- Reject if the P value is $\leq$ 0.05 and conclude that either X_4 or X_5 or both contain additional information useful for predicting Y in a linear model that also includes X_1, X_2, and X_3

Examples

- Predict bone density using age, weight and height; does diet add any useful information?
- Predict GPA using 3 HS grade variables; do SAT scores add any useful information?

Examples (2)

- Predict yield of an industrial process using temperature and pH; does the supplier of the raw material (categorical) add any useful information?

Extra SS Special Cases

- Compare models that differ by one explanatory variable, F(1,n-p)=t^2(n-p)
- t test for the hypothesis b_i=0 is equivalent to the general linear test based on

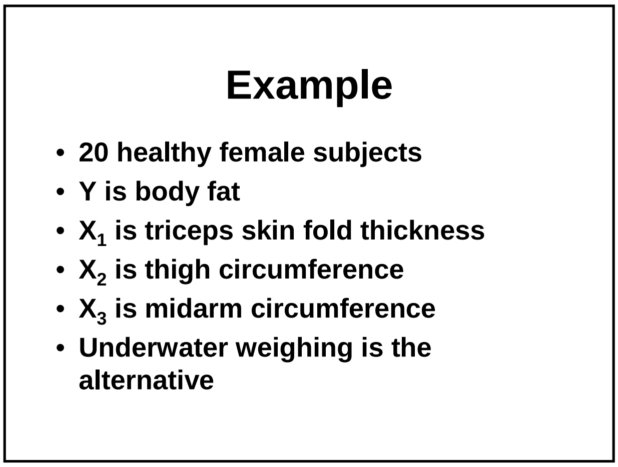
  SSM(X_i|X_1,..., X_{i-1}, X_{i+1} ,..., X_{p-1}) – Type II SS in R

Type I SS in R (default)

- *Add* one variable at a time
  - *SSM* (X_1)
  - *SSM* (X_2 | X_1)
  - *SSM* (X_3 |X_1, X_2)
  - *SSM* (X_4 |X_1, X_2, X_3)

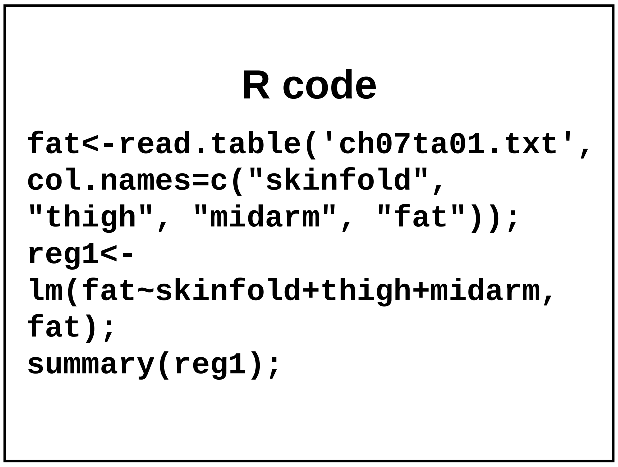
One Variable added

- *SSM* (X_1), *SSM* (X_2 | X_1), *SSM* (X_3 |X_1, X_2), *SSM* (X_4 |X_1, X_2, X_3)
- Df = 1 for each of these
- F = (SS/1) / MSE(full) ~ F(1, n-k)
- This is Type I SS in R
- *SSM* (X_1) +*SSM* (X_2 | X_1) + *SSM* (X_3 |X_1, X_2) + *SSM* (X_4 |X_1, X_2, X_3) =SSM(X_1, X_2, X_3, X_4)



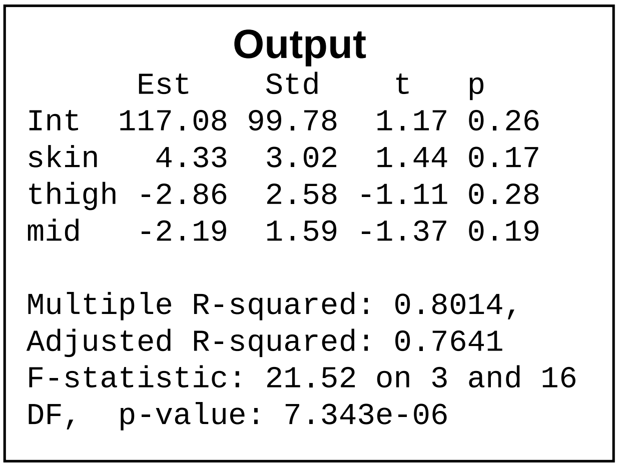
Example

- 20 healthy female subjects
- Y is body fat
- X_1 is triceps skin fold thickness
- X_2 is thigh circumference
- X_3 is midarm circumference
- Underwater weighing is the alternative



R code

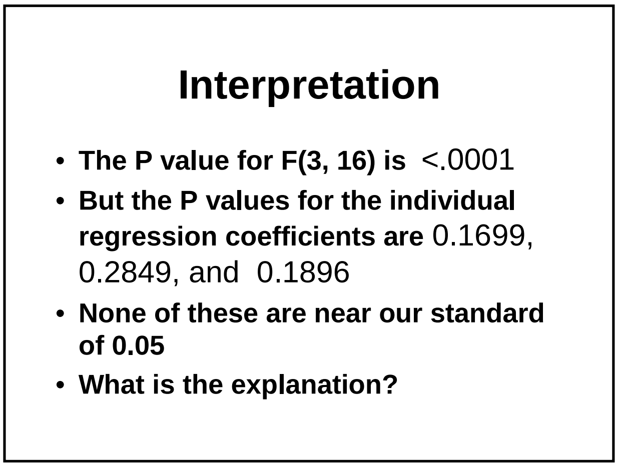
```
fat<-read.table('ch07ta01.txt',
col.names=c("skinfold",
"thigh", "midarm", "fat"));
reg1<-
lm(fat~skinfold+thigh+midarm,
fat);
summary(reg1);
```



Output

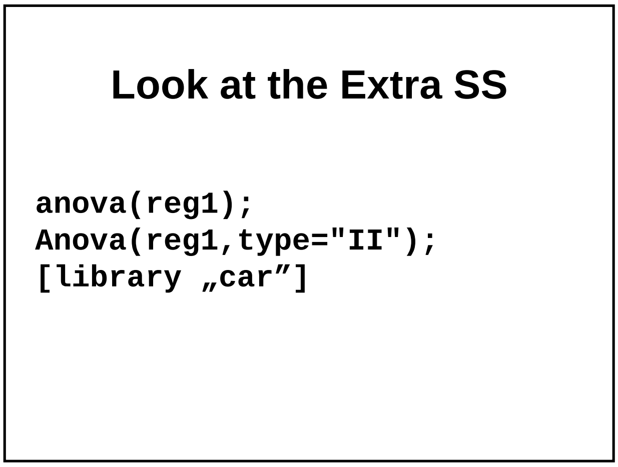
```
      Est    Std    t    p
Int  117.08 99.78  1.17 0.26
skin   4.33  3.02  1.44 0.17
thigh -2.86  2.58 -1.11 0.28
mid   -2.19  1.59 -1.37 0.19

Multiple R-squared: 0.8014,
Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16
DF,  p-value: 7.343e-06
```



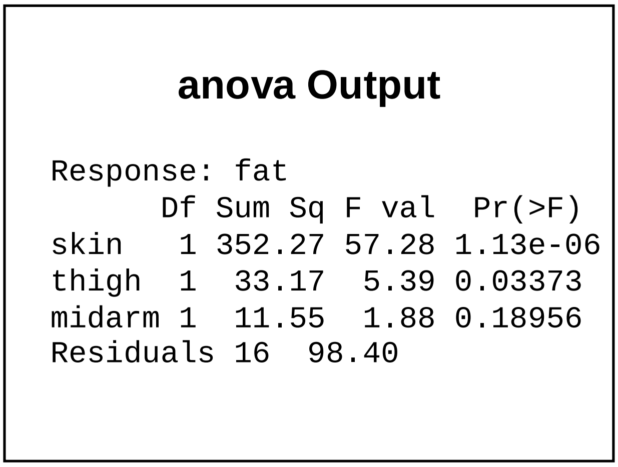
Interpretation

- **The P value for F(3, 16) is** <.0001
- **But the P values for the individual regression coefficients are** 0.1699, 0.2849, and 0.1896
- **None of these are near our standard of 0.05**
- **What is the explanation?**



Look at the Extra SS

```
anova(reg1);
Anova(reg1,type="II");
[library „car”]
```



anova Output

```
Response: fat
      Df Sum Sq F val  Pr(>F)
skin   1 352.27 57.28 1.13e-06
thigh  1  33.17  5.39 0.03373
midarm 1  11.55  1.88 0.18956
Residuals 16  98.40
```

- Anova Table (Type II tests)
- Response: fat
-             Sum   Df F value Pr(>F)
- skinfold 12.705 1 2.0657 0.1699
- thigh   7.529 1 1.2242 0.2849
- midarm   11.546 1 1.8773 0.1896
- Residuals 98.405 16

Interpretation

- Fact: the Type I and Type II SS are very different
- If we reorder the variables in the model statement we will get
  - Different Type I SS
  - The same Type II SS

Run additional models

- Rerun with skinfold as the explanatory variable

```
reg2<-lm(fat~skinfold, fat);
summary(reg2);
```

Output

```
      Est  Std    t    p
Int  -1.50 3.32 -0.45 0.66
skin  0.86 0.13  6.66 3.02e-06
```

Testing for remaining variables

```
anova(reg2,reg1);
Analysis of Variance Table

Mod 1: fat ~skinfold
Mod 2: fat ~skinfold+thigh+midarm
   dfE RSS   Df SS   F     Pr(>F)
1  18 143.12
2  16  98.40 2 44.72 3.64 0.04995
```

Other uses

- GL tests can be used to perform a significance test for any hypothesis involving a linear combination of the regression coefficients
- Examples

H_0: $\beta_4 = \beta_5$ (model: I(x4+x5)+x1+…)

H_0: $\beta_4 - 3\beta_5 = 12$ [y12*x4~I(3*x4+x5)+…]

Partial correlations

- Measures the strength of a linear relation between two variables taking into account other variables
- Procedure to find partial correlation
  - Predict Y with conditioning on other X's
  - Predict X_i with conditioning on other X's
  - Find correlation between the two sets of residuals

Coefficients of Partial Determination

- Measures the percentage reduction in SSE due to one explanatory variable when all the others are already included in the model

$$r^2{}_i = \frac{SSM(X_i \mid X_1,\dots,X_{i-1},X_{i+1},\dots,X_{p-1})}{SSE(X_1,\dots,X_{i-1},X_{i+1},\dots,X_{p-1})}$$

$$= \frac{\mathrm{SS_i}II}{SSE(F)+SS_iII}$$

Standardized Regression Model

- Can help reduce round off errors in calculations
- Puts regression coefficients in common units
- Units for the usual coefficients are units for Y divided by units for X

Standardized Regression Model (2)

- Standardized can be obtained from the usual ones by multiplying by the ratio of the standard deviation of X to the standard deviation of Y
- Interpretation is that a one sd increase in X corresponds to a 'standardized beta' increase in Y

Standardized Regression Model (3)

- Y = … + βX + …
- = … + $\beta(s_X/s_Y)(s_Y/s_X)X$ + …
- = … + $(\beta(s_X/s_Y))\ ((s_Y/s_X)X)$ + …
- = … + $(\beta(s_X/s_Y))\ (s_Y)\ (X/s_X)$ + …

Standardized Regression Model (4)

- Standardize Y and all X's (subtract mean and divide by standard deviation)
- The regression coefficients for variables transformed in this way are the standardized regression coefficients

R code

```
library(QuantPsyc)
lm.beta(reg1)

skinfold     thigh    midarm
 4.263705 -2.928701 -1.561417
```

Multicollinearity

- Numerical analysis problem is that the matrix X'X is close to singular and is therefore difficult to invert accurately
- Statistical problem is that there is too much correlation among the explanatory variables and it is therefore difficult to determine the regression coefficients

Multicollinearity (2)

- Solve the statistical problem and the numerical problem will also be solved
  - We want to refine a model that has redundancy in the explanatory variables even if X'X can be inverted without difficulty

Multicollinearity (3)

- Extremes cases can help us to understand the problem
  - if all columns in X matrix are uncorrelated, Type I SS and Type II SS will be the same, i.e, the contribution of each explanatory variable to the model will be the same whether or not the other explanatory variables are in the model

Multicollinearity (4)

- Extremes cases can help us to understand the problem
  - if there is a linear combination of the explanatory variables that is a constant (e.g. $X_1 = X_2$ $(X_1 - X_2 = 0)$), then the Type II SS for the X's involved will be zero

An example

```
cs<-read.table('csdata.dat',
col.names=c("id", "gpa", "hsm",
"hss", "hse", "satm",
"satv", "gen"));
cs$sat<-cs$satm+cs$satv;
reg3<-lm(gpa~sat+satm+satv,cs);
summary(reg3);
```

Output

```
Coeff: (1 not defined because
         of singularities)
    Estimate  Std.     t   p
Int   1.29    0.38     3.43 0.0007
sat -2.5e-05 6.2e-04 -0.04 0.9
satm 2.4e-03 1.1e-03 2.10 0.04
satv  NA       NA      NA  NA
F-stat: 7.476 on 2 and 221 DF,
p-value: 0.0007218
```

Extent of multicollinearity

- Our CS example had one explanatory variable equal to a linear combination of other explanatory variables
- This is the most extreme case of multicollinearity and is detected by statistical software because (X'X) does not have an inverse
- We are concerned with cases less extreme

Effects of multicollinearity

- Regression coefficients are not well estimated and may be meaningless
- Similarly for standard errors of these estimates
- Type I SS and Type II SS will differ
- R^2 and predicted values are usually ok

Two separate problems

- Numerical accuracy
  - (X'X) is difficult to invert
  - Need good software
- Statistical problem
  - Results are difficult to interpret
  - Need a better model

Polynomial regression

- We can do linear, quadratic, cubic, etc. by defining squares, cubes, etc. in a data step and using these as predictors in a multiple regression
- We can do this with more than one explanatory variable
- When we do this we generally create a multicollinearity problem

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

Input and check the data

```
cell<-read.table('ch08ta01.txt',
col.names=c("cycles", "chrate",
"temp"));
cell1<-cell;
```

```
      cycles chrate temp
 1     150    0.6   10
 2      86    1.0   10
 3      49    1.4   10
 4     288    0.6   20
 5     157    1.0   20
 6     131    1.0   20
 7     184    1.0   20
 8     109    1.4   20
 9     279    0.6   30
 10    235    1.0   30
 11    224    1.4   30
```

Create the new variables and run the regression

```
cell1$chr2<-cell1$chrate^2;
cell1$tm2<-cell1$temp^2;
cell1$chrtm<-
cell1$chrate*cell1$temp;
reg4<-lm(cycles~chrate+temp+chr2+
tm2+chrtm,cell1);
summary(reg4);
```

```
        Est     Std      t    p
Int     337.72 149.96   2.25 0.07
Chrate -539.52 268.86  -2.01 0.10
temp      8.92   9.18   0.97 0.38
chr2    171.22 127.13   1.35 0.24
tm2      -0.11   0.20  -0.52 0.62
chrtm     2.87   4.05   0.71 0.51
Multiple R-squared: 0.9135,
F-statistic: 10.57 on 5 and 5 DF,
  p-value: 0.01086
```

Conclusion

- **We have a multicollinearity problem**
- **Lets look at the correlations (use cor(cell))**
- **There are some very high correlations**
  - **r(chrate,chr2) =** 0.99103
  - **r(temp,tm2) =** 0.98609

A remedy

- **We can remove the correlation between explanatory variables and their squares**
- **Center (subtract the mean) before squaring**

```
cell2<-scale(cell);
•              cycles    chrate      temp
•   [1,] -0.2825953 -1.290994 -1.290994
•   [2,] -1.1046906  0.000000 -1.290994
•   [3,] -1.5799644  1.290994 -1.290994
•   [4,]  1.4900477 -1.290994  0.000000
•   [5,] -0.1926786  0.000000  0.000000
•   [6,] -0.5266548  0.000000  0.000000
•   [7,]  0.1541429  0.000000  0.000000
•   [8,] -0.8092501  1.290994  0.000000
•   [9,]  1.3744406 -1.290994  1.290994
•  [10,]  0.8092501  0.000000  1.290994
•  [11,]  0.6679524  1.290994  1.290994
```

Recompute squares and cross product

```
chr2<-cell2[,2]^2;
tm2<-cell2[,3]^2;
chrtm<-cell2[,2]*cell2[,3];
reg5<-lm(cell2[,1]~cell2[,2]+
cell2[,3]+chr2+tm2+chrtm);
summary(reg5);
```

```
•                   Estimate  Std. Error t value Pr(>|t|)
•  (Intercept) -0.11764    0.21333  -0.551  0.60508
•  cell2[, 2]  -0.55553     0.13150  -4.224  0.00829 **
•  cell2[, 3]   0.75122     0.13150   5.712  0.00230 **
•  chr2         0.21114     0.15676   1.347  0.23586
•  tm2         -0.08174     0.15676  -0.521  0.62435
•  chrtm        0.08863     0.12476   0.710  0.50918
•  ---
•  Multiple R-squared: 0.9135,
•  F-statistic: 10.57 on 5 and 5 DF,  p-value: 0.01086
```

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

- 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

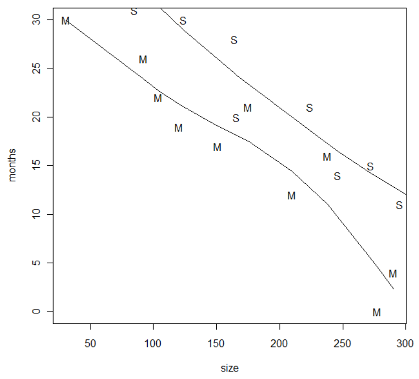
Example

- 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

```
funds<-read.table('ch08ta02.txt',
col.names=c("months", "size", "stock"));
v1<-funds[(funds$stock==0),];
v2<-funds[(funds$stock==1),];
plot(months~size,pch="M",v1);
u1<-order(v1$size);
v1<-v1[u1,];
s<-smooth.spline(v1$size,v1$months,
spar=0.5);
lines(s);
points(months~size,pch="S",v2);
u2<-order(v2$size);
v2<-v2[u2,];
s<-smooth.spline(v2$size,v2$months,
spar=0.5);
lines(s);
```



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 ? (GL test)
- funds$sizestock<-funds$size*funds$stock;
- reg1<-lm(months~size+stock+sizestock);
- reg2<-lm(months~size);
- anova(reg2,reg1);

Output

```
Model 1: months ~ size
Model 2: months ~ size + stock
+ sizestock

   RSS     Df   SS      F     Pr(>F)
1  492.63
2  176.38   2 316.25 14.34 0.0003
```

Output (3)
How are they different ?

```
summary(reg1)
             t value Pr(>|t|)
(Intercept) 13.864 2.47e-10 ***
size        -7.779 7.97e-07 ***
stock        2.225   0.0408 *
sizestock   -0.023   0.9821

Multiple R-squared: 0.8951
```

Two parallel lines

```
reg3<-lm(months~size+stock,
funds);
summary(reg3);
```

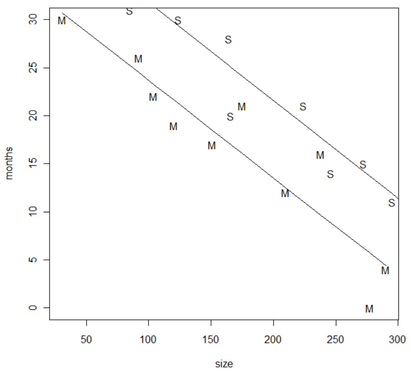
Output (3)

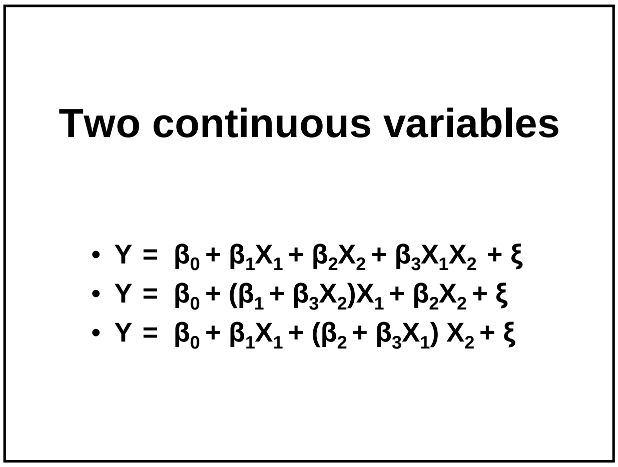
```
        Error t value
Pr(>|t|)
Int  33.87 1.81 18.675 9.15e-13
Size -0.10 0.01 -11.443 2.07e-9
stock 8.05 1.46   5.521 3.74e-5
Multiple R-squared: 0.8951,
```

Int for stock firms is 33.87+8.05 = 41.92

Plot the two lines

```
s<-smooth.spline(v1$size,
v1$months, spar=1);
lines(s);
s<-smooth.spline(v1$size,
v1$months, spar=1);
```





Two continuous variables

- $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \xi$
- $Y = \beta_0 + (\beta_1 + \beta_3X_2)X_1 + \beta_2X_2 + \xi$
- $Y = \beta_0 + \beta_1X_1 + (\beta_2 + \beta_3X_1) X_2 + \xi$