STOR 455
STATISTICAL METHODS I

Jan Hannig

Model building

- Often a large number of predictors are available.
- Question – which ones to use
  - If important predictors are omitted, we clearly do not have a good prediction
  - Why not use all of the predictors?
  - If predictors that are not important are included, variance is increased and overfitting can occur (overfitting = model fits really well for our data but does not generalize to other data sets.)

Examples

- Predict bone density using age, weight and height; does diet add any useful information?
- Predict faculty salaries using highest degree, rank, time in rank, and department; does race (gender) make any difference?

Examples (2)

- Predict GPA using 3 HS grade variables; do SAT scores add any useful information?
- Predict yield of an industrial process using temperature and pH; does the supplier of the raw material (categorical) add any useful information?

Extra Sums of Squares

- Reduction in SSE when one or more variables added to the model
- Equivalent to increase in SSR as SSR+SSE=SSTO
- Useful for model comparison

Extra SS (2)

- Models we compare are nested. i.e., 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
- First includes all Xs of second

Extra SS (3)

- There is an F test to compare the two models
  - Test the null hypothesis that the regression coefficients for the *extra* variables are all zero
  - Ex: 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

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
- $SSR(X_4, X_5 \mid X_1, X_2, X_3)$ is the difference

$$SSR(X_4, X_5 \mid X_1, X_2, X_3) =$$
$$= SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4, X_5)$$
$$= SSR(X_1, X_2, X_3, X_4, X_5) - SSR(X_1, X_2, X_3)$$

Extra SS (4)

- Df for the F statistic: the number of *extra* variables and the df for the model with larger number of explanatory variables
- Ex: 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 (n-4)-(n-6)=2
- Denominator df is n-6 = 94

F test

- Numerator is $MSR(X_4, X_5 \mid X_1, X_2, X_3)$
- Denominator is $MSE(X_1, X_2, X_3, X_4, X_5)$
- $F \sim F(2, n-6)$
- Reject if the P value is small 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

Body Fat 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

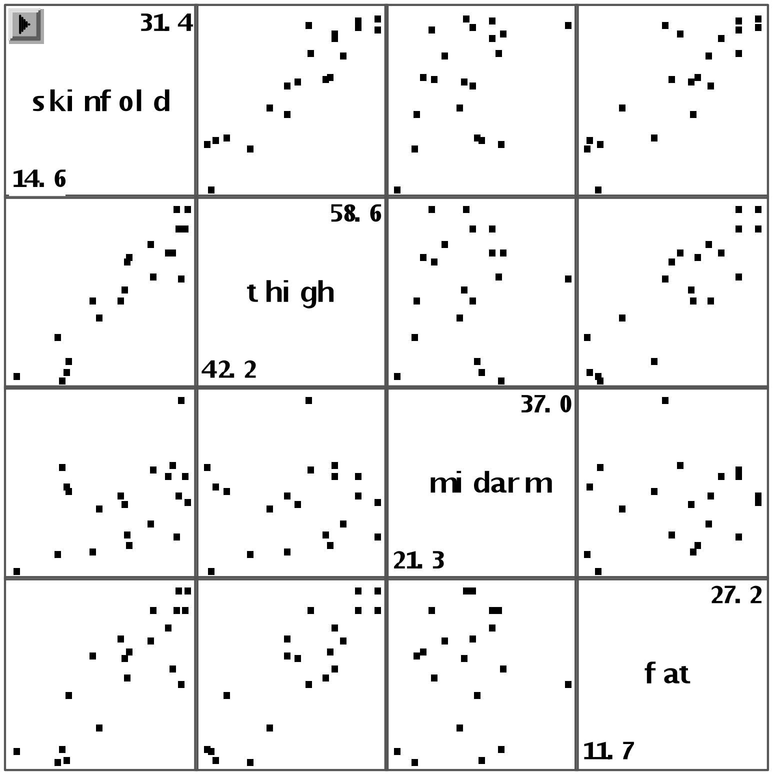
Do it in SAS

```
data fat;
  infile 'T:\...\Ch07ta01.txt';
  input skinfold thigh midarm fat;
proc print data=fat;
run;
```

Obs	skinfold	thigh	midarm	fat
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
...				
18	30.2	58.6	24.6	25.4
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

Do it in SAS

```
%include "T:\...
  \scatter.sas";
%scatter(data =
  fat, var
  =skinfold thigh
  midarm fat);
```



SAS Type I SS

- Compare models that differ by one explanatory variable
- *Add* one variable at a time
  - *SSR* (X_1)
  - *SSR* $(X_2 \mid X_1)$
  - *SSR* $(X_3 \mid X_1, X_2)$
  - *SSR* $(X_4 \mid X_1, X_2, X_3)$
- Order matters!

Type I SS

- *SSR* (X_1), *SSR* $(X_2 \mid X_1)$, *SSR* $(X_3 \mid X_1, X_2)$, *SSR* $(X_4 \mid X_1, X_2, X_3)$
- Df = 1 for each of these
- F = (SS/1) / MSE(full) ~ F(1, n-p)
- $F(1,n\text{-}p)=t^2(n\text{-}p)$
- Test results depend on the order the variable is added to the model, but

  SSR=*SSR* (X_1)+*SSR* $(X_2 \mid X_1)$+*SSR* $(X_3 \mid X_1, X_2)$+…

SAS Type II SS

- Different from Type I SS unless X_i independent
- Order of the variables does not matter

  SSR $(X_1 \mid X_2, X_3, X_4)$
  SSR $(X_2 \mid X_1, X_3, X_4)$
  SSR $(X_3 \mid X_1, X_2, X_4)$
  SSR $(X_4 \mid X_1, X_2, X_3)$
- The type II sum of squares do not add to anything interesting

Type II SS

- Df = 1 for each of these
- $F = (SS/1) / MSE(full) \sim F(1, n\text{-}p)$
- $F(1,n\text{-}p)=t^2(n\text{-}p)$
- This test is equivalent to the t-test for each explanatory variable

Do it in SAS

```
*Type I and Type II sums of squares;
proc reg data=fat;
  model fat=skinfold thigh midarm /ss1 ss2;
run;
```

Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	R-Square	0.8014
Dependent Mean	20.19500	Adj R-Sq	0.7641
Coeff Var	12.28017		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	117.08469	99.78240	1.17	0.2578	8156.76050	8.46816
skinfold	1	4.33409	3.01551	1.44	0.1699	352.26980	12.70489
thigh	1	-2.85685	2.58202	-1.11	0.2849	33.16891	7.52928
midarm	1	-2.18606	1.59550	-1.37	0.1896	11.54590	11.54590

F-test in SAS

```
*F-test;
proc reg data=fat;
   model fat=skinfold thigh midarm;
   test1: test thigh, midarm;
   test2: test thigh=1;
run;
```

Do it in SAS

Test test1 Results for Dependent Variable fat

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	22.35741	3.64	0.0500
Denominator	16	6.15031		

Test test2 Results for Dependent Variable fat

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	13.72284	2.23	0.1547
Denominator	16	6.15031		

F-test in SAS

```
*F-test;
proc reg data=fat;
  model fat=skinfold midarm;
     test skinfold=1;
run;
```

Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	389.45533	194.72767	31.25	<.0001
Error	17	105.93417	6.23142		
Corrected Total	19	495.38950			

Root MSE	2.49628	R-Square	0.7862
Dependent Mean	20.19500	Adj R-Sq	0.7610
Coeff Var	12.36089		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.79163	4.48829	1.51	0.1486
skinfold	1	1.00058	0.12823	7.80	<.0001
midarm	1	-0.43144	0.17662	-2.44	0.0258

Do it in SAS

Model: MODEL1

Test 1 Results for Dependent Variable fat

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.00012965	0.00	0.9964
Denominator	17	6.23142		

Partial correlations

- Measure the strength of a linear relation between two variables taking into account (or conditioning on) other variables
- *Coefficient of partial determination:* squared partial correlation
- $r^2_{Y1,234}$=*SSR* (X_1 |X_2, X_3, X_4)/SSE(X_2, X_3, X_4)

Partial correlations (2)

- Equivalent to
  - Predict Y with conditioning X's
  - Predict X_i with conditioning X's
  - Find (squared) correlation between the two sets of residuals
- SAS option /PCORR1 /PCORR2

Do it in SAS

```
*Partial correlations;
proc reg data=fat;
   model fat=skinfold thigh midarm / pcorr1 pcorr2;
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type I	Squared Partial Corr Type II
Intercept	1	117.08469	99.78240	1.17	0.2578	.	.
skinfold	1	4.33409	3.01551	1.44	0.1699	0.71110	0.11435
thigh	1	-2.85685	2.58202	-1.11	0.2849	0.23176	0.07108
midarm	1	-2.18606	1.59550	-1.37	0.1896	0.10501	0.10501