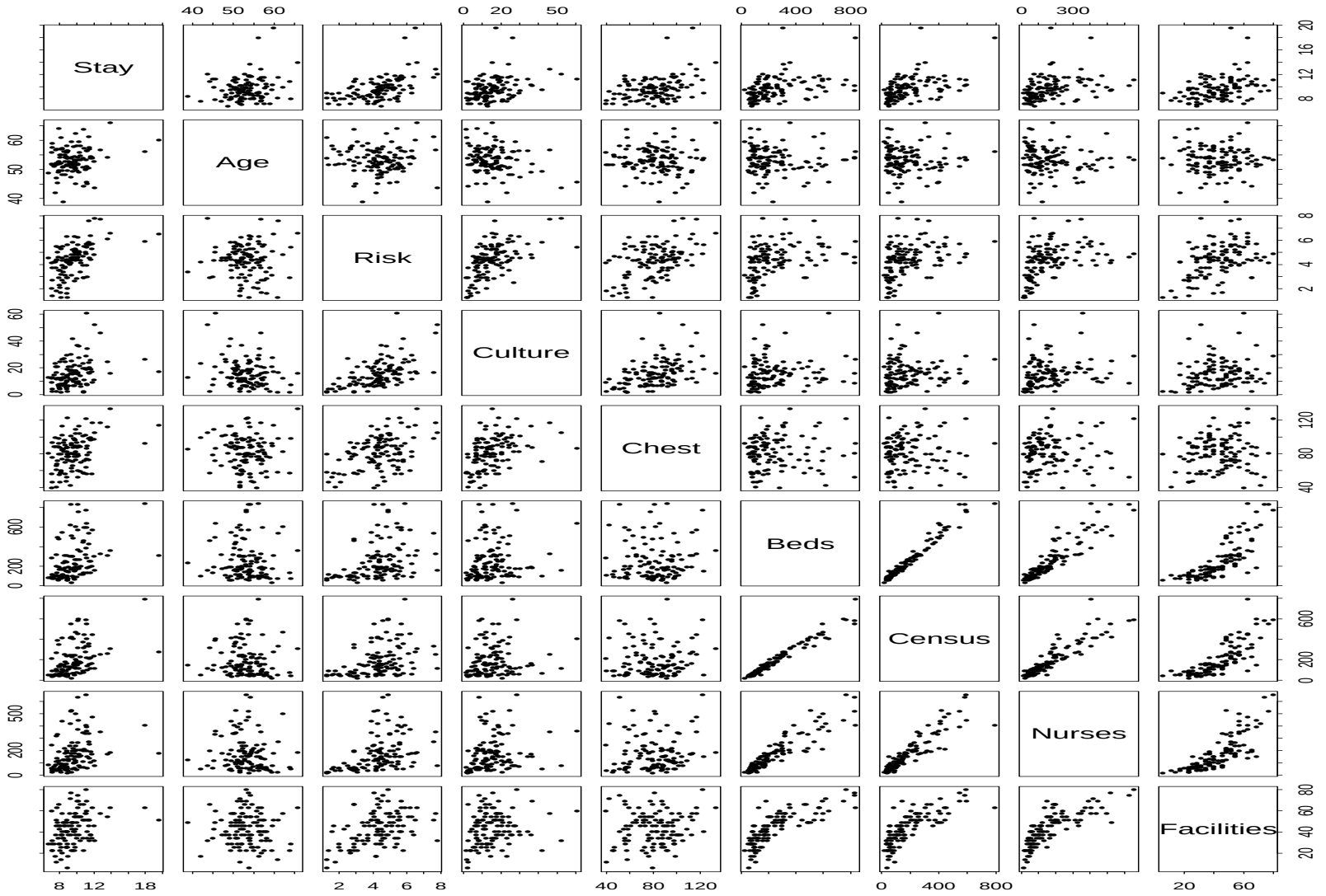
SENIC data example

Sample of 113 hospitals; observational study. Variables measured:

- Average length of stay of patients in days
- Average age of patients.
- Probability of acquiring infection in hospital. (How?)
- Culturing ratio: 100 times the ratio (Cultures performed) divided by (number of patients with no infection).
- Chest X-ray ratio defined similarly.
- Number of beds.
- Medical school affiliation (A dichotomous, Yes or no, variable).
- Geographic region (in the US) – NE, NC, S or W.
- Number of patients.
- Number of nurses.
- Available facilities (available at the given hospital).

Pairwise Scatter Plots



Comments

As expected, several of the variables are quite highly correlated. Here is the correlation matrix:

	Stay	Age	Risk	Culture	Chest	Beds	Census	Nurses	Facilities
Stay	1.00	0.19	0.53	0.33	0.38	-0.49	0.47	0.34	0.36
Age	0.19	1.00	0.00	-0.23	-0.02	-0.02	-0.05	-0.08	-0.04
Risk	0.53	0.00	1.00	0.56	0.45	-0.19	0.38	0.39	0.41
Culture	0.33	-0.23	0.56	1.00	0.42	-0.31	0.14	0.20	0.19
Chest	0.38	-0.02	0.45	0.42	1.00	-0.30	0.06	0.08	0.11
Beds	0.41	-0.06	0.36	0.14	0.05	-0.11	0.98	0.92	0.79
Census	0.47	-0.05	0.38	0.14	0.06	-0.15	1.00	0.91	0.78
Nurses	0.34	-0.08	0.39	0.20	0.08	-0.11	0.91	1.00	0.78
Facilities	0.36	-0.04	0.41	0.19	0.11	-0.21	0.78	0.78	1.00

- Fit several models and discuss interpretation of the coefficients.
- Think about how the variables should influence Risk.
- Risk of infection should increase with length of stay.
- Data provides only ecological correlations — risk for a whole group of patients is being related to average stay.
- Nothing in data indicates clearly whether or not patients with long stays are actually getting infected more often.
- But: seems reasonable that if we could hold other features of the hospital environment constant then hospitals with long average stays would expose their patients to more risk, i.e., would have a higher probability of infection.
- Similar remarks should be made about all discussion below.

More hypotheses

- Age expected to be positively correlated with risk of infection — presumably, older patients are more susceptible.
- Chest and Culture are measures of how hard the hospital looks for otherwise unsuspected infection.
- If you look harder you should find more so that the coefficients of these variables in any regression model would be expected to be positive.
- Beds, Census and Nurses all measure the size of the hospital.
- Not clear if a big hospital should put a patient at risk of infection or not.
- However, I point out that the relations between these three variables might make a difference.
- A hospital with high Census compared to Beds may be overcrowded and so more likely to support infections.

More Hypotheses

- A hospital with lots of Nurses relative to patients might be expected to be kept in better condition and so lower the risk.
- Facilities seems to measure the sophistication of medical treatment at the hospital.
- This might be positive or it might suggest more exotic diseases among the patient population so I can't guess intelligently about the direction of the effect.
- These hypotheses should be generated in advance of collecting data;
- They can then be checked formally by multiple regression.

First Models

We begin by fitting a model using all the continuous predictors, that is ignoring only SCHOOL and REGION. Here is Splus code and results.

```
> fit.full <- lm(Risk ~ Stay + Age + Culture
  + Chest + Beds + Census + Nurses + Facilities,
  data=scenic)
> summary(fit.full)

Call: lm(formula = Risk ~ Stay + Age + Culture
   + Chest + Beds + Census + Nurses + Facilities,
   data = scenic)
Residuals:
    Min      1Q  Median      3Q   Max
 -2.222 -0.5918 0.01833 0.5453 2.556
```

S-Plus Output Continued

```
Coefficients:
               Value Std. Error t value Pr(>|t|)
(Intercept) -0.7473  1.2076    -0.6188  0.5374
      Stay  0.1769  0.0691     2.5621  0.0118
       Age  0.0162  0.0223     0.7285  0.4679
   Culture  0.0470  0.0107     4.3719  0.0000
     Chest  0.0120  0.0055     2.1943  0.0304
      Beds -0.0014  0.0027    -0.5340  0.5945
    Census  0.0007  0.0035     0.2097  0.8343
    Nurses  0.0019  0.0018     1.0873  0.2794
Facilities  0.0163  0.0102     1.5978  0.1131

Residual standard error: 0.959 on 104 deg. of fr.
Multiple R-Squared: 0.5251
F-statistic: 14.37 on 8 and 104  deg. of fr.,
             the p-value is 6.117e-14
```

S-Plus Output Continued

```
Correlation of Coefficients:
        Intercept  Stay    Age    Culture  Chest   Beds  Census  Nurses
    Stay -0.0146

     Age -0.8622  -0.3347

 Culture -0.1606  -0.2930  0.3042

   Chest -0.1914  -0.3039  0.0281 -0.2911

    Beds -0.0377   0.2726 -0.0775 -0.0551  0.0077

  Census  0.0536  -0.4625  0.1369  0.1530  0.0657 -0.8766

  Nurses  0.0041   0.2486 -0.0410 -0.1935 -0.0657 -0.1681 -0.2265

 Facilit -0.1544  -0.0475 -0.0232 -0.0363 -0.0613 -0.2009  0.0670 -0.2229
```

Discussion

- Output suggests that Stay, Culture and Chest are important predictors but none of the others are.
- So we fit next the model retaining only these 3 predictors.
- Warning: strategy being followed now is flawed.

Second Model

```
> fit.1 <- lm( Risk ~ Stay + Culture + Chest,
                                        data=scenic)
> summary(fit.1)

Residuals:
    Min      1Q   Median     3Q    Max
 -2.181 -0.7678 -0.04002 0.696 2.594

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 0.3092 0.5425     0.5700  0.5698
      Stay 0.2450 0.0540     4.5349  0.0000
   Culture 0.0494 0.0103     4.8008  0.0000
     Chest 0.0110 0.0056     1.9839  0.0498
```

S-Plus Output Continued

```
Residual standard error: 0.99 on 109 deg of fr
Multiple R-Squared: 0.4696
F-statistic: 32.16 on 3 and 109  deg. of fr.,
             the p-value is 5.662e-15

Correlation of Coefficients:
         (Intercept)    Stay Culture
    Stay -0.6633
Culture  0.1766     -0.1963
  Chest -0.4611     -0.2848 -0.3435
```

Discussion

- Compare these models in Splus by carrying out an extra Sum of Squares F-test.
- I have edited the output to make it fit – changing the columns labelled Terms to the shorthand FULL and REDUCED

ANOVA in S-Plus

```
> anova(fit.1,fit.full)
Analysis of Variance Table

Response: Risk

 Terms   Res Df   RSS     Test Df   SS        F     Pr(F)
REDUCED 109    106.821
FULL    104     95.640    5     11.181 2.4316 0.03973
```

Discussion

- Test of hypothesis of no influence of the 5 extra variables suggests that you cannot delete them all
- $P = 0.03973$ is marginally significant.
- Notice, though, individual t-tests for the 5 individual coefficients are not significant.
- To see what can happen consider adding each of the three variables which measure size

```
> fit.b <-  lm(Risk ~ Stay + Culture
   + Chest + Beds , data = scenic)
> fit.c <- lm(Risk ~ Stay + Culture
   + Chest + Census, data = scenic)
> fit.n <- lm(Risk ~ Stay + Culture
   + Chest + Nurses, data = scenic)
> summary(fit.b)

Call: lm(formula = Risk ~ Stay + Culture
              + Chest + Beds, data = scenic)
Residuals:
    Min      1Q  Median   3Q   Max
 -1.993 -0.7365 0.05695 0.66 2.291
```

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 0.4149 0.5309     0.7816  0.4362
      Stay 0.1845 0.0578     3.1940  0.0018
   Culture 0.0480 0.0101     4.7701  0.0000
     Chest 0.0130 0.0055     2.3761  0.0193
      Beds 0.0013 0.0005     2.5516  0.0121

Residual standard error: 0.9658 on 108 df
Multiple R-Squared: 0.4997
F-statistic: 26.97 on 4 and 108 df,
  the p-value is 1.554e-15
```

```
Correlation of Coefficients:
        (Intercept)    Stay Culture   Chest
   Stay -0.6351
Culture  0.1714     -0.1558
  Chest -0.4439     -0.3155 -0.3475
   Beds  0.0780     -0.4099 -0.0560  0.1424
> summary(fit.c)

Call: lm(formula = Risk ~ Stay + Culture
          + Chest + Census, data = scenic)
Residuals:
    Min      1Q  Median      3Q   Max
 -1.984 -0.7584 0.07387 0.6545 2.447
```

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 0.5233 0.5357     0.9770  0.3307
      Stay 0.1719 0.0599     2.8694  0.0049
   Culture 0.0484 0.0101     4.8199  0.0000
     Chest 0.0132 0.0055     2.3952  0.0183
    Census 0.0017 0.0007     2.5656  0.0117

Residual standard error: 0.9655 on 108 df
Multiple R-Squared: 0.5
F-statistic: 27 on 4 and 108 degrees of
        freedom, the p-value is 1.554e-15
```

```
Correlation of Coefficients:
        (Intercept)    Stay Culture   Chest
   Stay -0.6504
Culture  0.1683     -0.1542
  Chest -0.4270     -0.3189 -0.3452
 Census  0.1558     -0.4757 -0.0384  0.1497
> summary(fit.n)

Call: lm(formula = Risk ~ Stay + Culture
       + Chest + Nurses, data = scenic)
Residuals:
    Min      1Q  Median      3Q   Max
 -1.958 -0.7093 0.02961 0.5473 2.453
```

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 0.3703 0.5241     0.7065  0.4814
      Stay 0.1936 0.0549     3.5263  0.0006
   Culture 0.0456 0.0100     4.5505  0.0000
     Chest 0.0127 0.0054     2.3480  0.0207
    Nurses 0.0021 0.0007     2.9949  0.0034

Residual standard error: 0.9556 on 108 df
Multiple R-Squared: 0.5102
F-statistic: 28.13 on 4 and 108 degrees
   of freedom, the p-value is 5.551e-16
```

```
Correlation of Coefficients:
        (Intercept)    Stay Culture   Chest
   Stay -0.6417
Culture  0.1700     -0.1450
  Chest -0.4544     -0.3008 -0.3519
 Nurses  0.0389     -0.3126 -0.1276  0.1013
```

Discussion

- Each of the measures of size is significant; that is, it appears that you should add each of them.
- Measures are *Multi-collinear*.
- get no additional predictive power out of including more than one of them in the model.
- Look what happens when I add both Census and Beds:

```
> fit.bc <- lm(Risk ~ Stay + Culture + Chest
    + Census + Beds, data = scenic)
> summary(fit.bc)

Call: lm(formula = Risk ~ Stay + Culture + Chest
     + Census + Beds, data = scenic)
Residuals:
    Min      1Q  Median      3Q   Max
 -1.986 -0.7498 0.07771 0.6563 2.386

Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept) 0.4841 0.5742     0.8431  0.4010
      Stay 0.1760 0.0637     2.7611  0.0068
   Culture 0.0483 0.0101     4.7610  0.0000
     Chest 0.0131 0.0055     2.3797  0.0191
    Census 0.0011 0.0034     0.3243  0.7463
      Beds 0.0005 0.0026     0.1956  0.8453
```

```
Residual standard error: 0.9699 on
            107 degrees of freedom
Multiple R-Squared: 0.5002
F-statistic: 21.42 on 5 and 107 degrees
    of freedom, the p-value is 8.549e-15

Correlation of Coefficients:
        (Intercept)    Stay Culture   Chest  Census
   Stay -0.6906
Culture  0.1887     -0.1749
  Chest -0.3926     -0.3080 -0.3417
 Census  0.3715     -0.4140  0.0810  0.0513
   Beds -0.3492      0.3301 -0.0906 -0.0215 -0.9794
```

Discussion

- Neither Beds nor Census significant.
- Reason for confusion is in matrix of correlations.
- Recall $Var(\hat{\beta}) = \sigma^2(X^T X)^{-1}$.
- To get $cor(\hat{\beta}_i, \hat{\beta}_j)$ take element ij and divide by square root of product of ith and jth diagonal entries.
- Standard deviation σ cancels out.
- Notice high negative correlation $cor(\hat{\beta}_{\text{Beds}}, \hat{\beta}_{\text{Census}})$.
- Splus code for an ANOVA table comparing the model with Beds and Census to the model without them:

```
> anova(fit.1,fit.bc)
Analysis of Variance Table

Response: Risk
```

Model	RSS	df	Extra SS	Extra df	F	*P*
Stay, Culture, Chest	109	106.821				
Add: Census, Beds	107	100.648	6.172	2	3.28	0.041

Interpretation: Must retain at least one measure of size, but don't need both.

Finally look what happens if we put in Beds and Nurses.

```
> fit.nb <- lm(Risk ~ Stay + Culture
      + Chest + Nurses + Beds, data = scenic)
> summary(fit.nb)

Call: lm(formula = Risk ~ Stay + Culture
    + Chest + Nurses + Beds, data = scenic)
Residuals:
    Min      1Q  Median      3Q   Max
 -1.959 -0.7159 0.05094 0.5114 2.513
```

```
Coefficients:
             Value Std. Error t value Pr(>|t|)
(Intercept)  0.3523  0.5290     0.6660  0.5069
      Stay   0.1998  0.0582     3.4312  0.0009
   Culture   0.0451  0.0102     4.4383  0.0000
     Chest   0.0125  0.0055     2.2807  0.0245
    Nurses   0.0026  0.0017     1.5528  0.1234
      Beds  -0.0004  0.0012    -0.3336  0.7394

Residual standard error: 0.9596 on 107 df
Multiple R-Squared: 0.5107
F-statistic: 22.34 on 5 and 107 degrees
   of freedom, the p-value is 2.776e-15
```

```
Correlation of Coefficients:
        (Intercept)  Stay    Culture  Chest  Nurses
   Stay -0.6371
Culture  0.1819    -0.1818
  Chest -0.4364    -0.3217 -0.3285
 Nurses -0.0763     0.1693 -0.1821 -0.0678
   Beds  0.1019    -0.3230  0.1422  0.1211 -0.9079
```

Notice particularly that Beds now has a negative coefficient (though not significantly so).

Summary

- Observational study — cannot interpret parameter estimates as measuring the amount by which the response would be expected to increase if you increased the corresponding variable by 1 unit.
- Trouble: in population, large values of Beds go with, usually, larger values of Census.
- So the coefficient of Beds measures something rather more like: how much would the response increase if I increased Beds by 1 unit and all the other covariates not in the model changed as you would expect from the relations in the variables seen in this population.
- Thus you can't really base a policy decision about building more beds in a hospital on the sign of the coefficient of Beds in a regression model like this.

SUMMARY

- OBSERVATIONAL STUDY.
- Coefficient of Nurses DOES NOT measure effect of hiring nurses.
- In population Hosps w/ more Nurses have different levels of other vars. Combination of differences influences Risk.
- Need experimental control of variables. Often too expensive.

The SENIC data set, continued

- STAY, CULTURE and CHEST are significant
- We must retain one of the three variables BED, NURSES and CENSUS which measure size of the hospital.
- These three variables are multi-collinear.
- Picking the variable of the three which produces the largest multiple R^2 we go with NURSES.
- Now we look at the question of adding further variables to that 4 covariate model.

```
> anova(fit.n,fit.full)
Analysis of Variance Table
Response: Risk
          Res          Test
Model     Df   RSS      Df  SS      F      Pr(F)
FULL     108 98.629
REDUCED  104 95.640   4 2.9895 0.8127 0.5198
```

This suggests we need not consider adding further variables.
However, we should examine diagnostics and consider the question of how variables are likely to influence RISK.
Suggestion: Transform other variables.
Define NURSE.RATIO = NURSES/CENSUS. Idea: large values indicate more intensive nursing care.
Define CROWDING = CENSUS/BEDS. Idea: large values indicate a crowded hospital.
Add these variables to the model.

```
> Nurse.Ratio <- scenic$Nurse/scenic$Census
> Crowding <- scenic$Census/scenic$Beds
> sc.ext <- data.frame(sc.ext, Nurse.Ration,Crowding)
> fit.l20 <- lm(Risk ~ Stay + Culture + Chest +
    Nurses + Crowding + Nurse.Ratio, data = sc.ext)
> summary(fit.l20)
Residuals:
    Min      1Q  Median      3Q   Max
 -2.036 -0.6102 0.01268 0.3956 2.798
Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) -1.2762  0.8753    -1.4581  0.1478
       Stay  0.2196  0.0594     3.6983  0.0003
    Culture  0.0424  0.0099     4.2740  0.0000
      Chest  0.0093  0.0055     1.7040  0.0913
     Nurses  0.0014  0.0007     1.9627  0.0523
   Crowding  1.4296  0.9455     1.5121  0.1335
Nurse.Ratio  0.8238  0.3298     2.4979  0.0140
```

```
Residual standard error: 0.9359 on 106 df
Multiple R-Squared: 0.5389
F-statistic: 20.65 on 6 and 106 df,
the p-value is 6.661e-16

Correlation of Coefficients:
        Int    Stay    Cult     Chest  Nurses Crowd
Stay  -0.3314
Cult   0.1738 -0.1725
Chest -0.1170 -0.3422 -0.3010
Nurse  0.3162 -0.2737 -0.0803  0.1608
Crowd -0.7108 -0.2136 -0.0321 -0.0605 -0.3032
N.Rat -0.6321  0.2561 -0.1365 -0.2548 -0.3056 0.3849
```

Conclusion: NURSE.RATIO is a useful predictor.
Can we discard CHEST, CROWDING? NURSES marginal but seems reasonable to keep this variable since we are keeping NURSE.RATIO.

```
fit.120.t <- lm(Risk ~ Stay + Culture
      + Nurse.Ratio + Nurses, data = sc.ext)
> summary(fit.120.t)
Residuals:
    Min      1Q  Median      3Q   Max
 -2.214 -0.6387 0.06483 0.5021 2.655

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) -0.0831  0.6092    -0.1365  0.8917
      Stay  0.2767  0.0549     5.0417  0.0000
   Culture  0.0482  0.0096     5.0311  0.0000
Nurse.Ratio 0.7695  0.2994     2.5701  0.0115
    Nurses  0.0016  0.0007     2.2607  0.0258
```

```
Residual standard error: 0.9511 on 108 df
Multiple R-Squared: 0.5149
F-statistic: 28.66 on 4 and 108 df,
         the p-value is 3.331e-16

Correlation of Coefficients:
            Int    Stay     Cult  N.Ratio
   Stay -0.8669
Culture  0.1569 -0.3317
N.Ratio -0.6468  0.3148 -0.2287
 Nurses  0.1916 -0.3356 -0.0521 -0.1851
> anova(fit.l20,fit.l20.t)
Analysis of Variance Table    Response: Risk
Model   Res df   ESS   test df   SS    F    P
FULL      106   92.852
REDUCED   108   97.689   2      4.84 2.76 0.068
```

Conclusion: Can discard CHEST, CROWDING but not NURSES.
Remaining Issues

- Diagnostics?
- Is this sequence of t, F tests a good way to select a model?
- Many tests done. Overall probability of no Type I or II errors?
- What about models we didn't try?
- Notice: CHEST significant at first then deleted after NURSES, NURSES.RATIO put in.
- Cause and effect: inference in an observational study is largely descriptive.

Cause and Effect

- Research question: do changes in variable X **cause** changes in Y?
- If so we could manipulate X and change Y.
- PROOF of cause and effect: hold all other important variables constant and try experimental units at various settings of X.
- Variables we don't know about or can't control are (probably) equalized between the different levels of X by randomly assigning units to the different values of X.
- Observational study: X cannot be controlled and other variables cannot be held constant.

Example

- Suppose men have generally higher values of both X and Y and women have generally lower values
- Suppose that among men there is no relation between X and Y.
- Suppose that among women there is no relation between X and Y.
- Overall correlation positive.
- Within sex no relation.
- Manipulating X leaves sex unchanged so Y is unaffected.
- Make comparison adjusting for sex by fitting separate lines in the two groups. Different intercepts **adjust** for sex.
- Multiple regression **adjusts** for the other covariates.
- But you can't adjust for variables you don't measure.

- Does decreasing nursing ratio lower the risk of nosocomial infection?
- Should you fire some nurses?
- No such deduction rigorously possible.
- 3rd variable not in list?

Adjusting for covariates

- The slope in a regression model corresponding to X measures the change expected in Y when X is changed by 1 unit and all the other variables in the regression are held constant.
- It is in this sense the regression method is used to *adjust* for the other covariates.
- Researchers say things like "Adjusted for Length of service and publication rate sex has no impact on salary of professors."
- See notes on "Experimental Design."