Overview of Topic IV

This topic will cover

- General Linear Tests (Chapter 7)
- Extra Sums of Squares (Chapter 7)
- Partial Correlations (Chapter 7)
- Multicollinearity (Chapter 7)
- Model Selection (Chapter 9)

General Linear Tests

These are a different way to look at the comparison of models. So far we have looked at comparing/selecting models based on ...

- model significance test and R^2 values
- t-tests for variables added last

These are good things to look at, but they are ineffective in cases ...

- explanatory variables work together in groups
- we want to test some hypotheses for some $\beta_i = b$ rather than $\beta_i = 0$ (for example, maybe we want to test $\mathrm{H}_0 : \beta_1 = 3, \beta_4 = 7$ against the alternative hypothesis that at least one of those is false)

General Linear Tests look at the difference between models

- in terms of SSE (unexplained SS)
- in terms of SSM (explained SS)

Because $SSM + SSE = SST$, these two comparisons are equivalent.

The models we compare are *hierarchical* in the sense that one (the full model) includes all of the explanatory variables of the other (the reduced model).

We can compare models with different explanatory variables. For example:

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 that the first model includes all X's of the second model.
- 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

$$\begin{aligned} \text{H}_0 &: \quad \beta_4 = \beta_5 = 0 \\ \text{H}_a &: \quad \beta_4 \text{ and } \beta_5 \text{ are not both } 0 \end{aligned}$$

F-test

The test statistic in general is

$$F = \frac{\left(SSE(R) - SSE(F)\right) / \left(df_E(R) - df_E(F)\right)}{SSE(F)/df_E(F)}$$

- Under the null hypothesis (reduced model) this statistic has an F distribution where the degrees of freedom are the number of *extra* variables and the df_E for the larger model.

- We reject if the p-value for this test is $p \leq 0.05$, and in that case conclude that at least one of the extra variables is useful for predicting Y in the linear model that already contains the variables in the reduced model.

Example

Suppose $n = 100$ and we are testing X_1, X_2, X_3, X_4, X_5 (full) vs X_1 , X_2 , X_3 (reduced). Our hypotheses are:

$$\begin{aligned} \text{H}_0 &: \quad \beta_4 = \beta_5 = 0 \\ \text{H}_a &: \quad \beta_4 \text{ and } \beta_5 \text{ are not both } 0 \end{aligned}$$

Since we are considering removing 2 variables (X_4 and X_5), the numerator df is 2. The denominator df is $n - 6 = 94$ (since $p = 6$ for the full model). We reject if the p-value ≤ 0.05 and in that case would conclude that either X_4 or X_5 or both contain additional information that is useful for predicting Y in a linear model that also includes X_1, X_2, and X_3.

Extra Sums of Squares

Notation for Extra SS

Example using 5 variables:

- $SSE(X_1, X_2, X_3, X_4, X_5)$ is the SSE for the *full* model, $SSE(F)$
- $SSE(X_1, X_2, X_3)$ is the SSE for a *reduced* model, $SSE(R)$.

- The ”extra sum of squares” for this comparison is denoted $SSM(X_4, X_5|X_1, X_2, X_3)$.
- This is the difference in the SSE’s:

$$\begin{aligned} SSM(X_4, X_5|X_1, X_2, X_3) &= SSE(R) - SSE(F) \\ &= SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4, X_5) \end{aligned}$$

- We call this an “*extra sum of squares*” because it is the extra amount of SS explained by the full model that is not explained by the reduced model. In this case it is the extra sums of squares explained by X_4 and X_5, given that X_1, X_2, and X_3 are already in the model.

For each model, $SST = SSM + SSE$, so $SSE = SST - SSM$, and SST is the same for both models. It follows that

$$\begin{aligned}
& SSM(X_4, X_5 | X_1, X_2, X_3) \\
&\quad = SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4, X_5) \\
&\quad = [SST - SSM(X_1, X_2, X_3)] - \\
&\qquad [SST - SSM(X_1, X_2, X_3, X_4, X_5)] \\
&\quad = SSM(X_1, X_2, X_3, X_4, X_5) - SSM(X_1, X_2, X_3) \\
&\quad = SSM(F) - SSM(R)
\end{aligned}$$

Thus, the extra sums of squares in the numerator of the test may also be written as the difference in the SSM's $(full - reduced)$.

Similarly, since $df_M + df_E = df_T$ for each model, we have

$$\begin{aligned} df_E(R) - df_E(F) &= [df_T - df_M(R)] - [df_T - df_M(F)] \\ &= df_M(F) - df_M(R) \end{aligned}$$

Thus the df in the numerator of the test may also be written as the difference in the df_M's $(full - reduced)$.
It doesn't matter whether you look at as the difference in the models or the difference in the errors, as long as you get the signs right.

F-test (General Linear Test)

- *Numerator*: $SSM(X_4, X_5|X_1, X_2, X_3)/2$ (where $2 =$ # variables before "$|$")
- *Denominator*: $MSE(X_1, X_2, X_3, X_4, X_5)$ (full model MSE)
- $F \sim F_{2,n-6}$
- Numerator $df = 2$ since two parameters fixed, i.e. difference in parameters $= full - reduced$.
- Denominator $df = n - 6$ since $p = 6$ for full model.

Reject if the p-value ≤ 0.05, and in that case would conclude that either X_4 or X_5 or both contain additional information that is 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?
- 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 (SAS)

Type II SS: Each variable given all others are in the model

- Compare models that differ by one explanatory variable, $F_{1,n-p} = t^2_{n-p}$ (i.e. ‘variable added last’ tests).
- SAS’s individual parameter t-tests are equivalent to the general linear test based on $SSM(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$
- These are called Type II SS (not related to Type II Error).
- Type II SS are the extra sums of squares for each variable, given that all other variables are in the model.
- Option `SS2` to `proc reg`.

Type I SS: Add one variable at a time in order

Consider 4 explanatory variables *in a particular order* and the extra sums of squares we get by adding one variable at a time, starting with the first one (SAS takes them in the order listed in the `model` statement), and testing each successive model with the one before it. . .

- $SSM(X_1)$
- $SSM(X_2|X_1)$
- $SSM(X_3|X_1, X_2)$
- $SSM(X_4|X_1, X_2, X_3)$

It is easy to work out from the definitions that these sum to the SSM for the full model:

$$\begin{aligned} SSM(X_1) + SSM(X_2|X_1) + SSM(X_3|X_1, X_2) + \\ SSM(X_4|X_1, X_2, X_3) \\ = SSM(X_1, X_2, X_3, X_4) \end{aligned}$$

For the tests,

- Numerator df is 1 for all of them (one parameter before "$|$").
- Denominator df depends on which model acts as the full one.
- $F = (SSR/1)/MSE(F) \sim F_{1,n-p}$
- These are called Type I SS (not related to Type I Error). Notice that the Type I SS depend on the order in which the variables are listed.
- Type I SS are the extra SS for each variable, given that *all previous* variables are in the model. ("previous" meaning having a lower subscript number)

KNNL Example page 256

- SAS code in `nknw260.sas`
- 20 healthy female subjects ages 25-34
- Y is fraction body fat
- X_1 is triceps skin fold thickness
- X_2 is thigh circumference
- X_3 is midarm circumference
- Goal is to find a good model based on these three easy measurements. Otherwise we will have to use underwater weight/density measurements (difficult and more expensive).

Step 1: Check the data and run `proc reg` on full model

```
data bodyfat;
   infile ’H:\System\Desktop\Ch07ta01.dat’;
   input skinfold thigh midarm fat;
proc print data=bodyfat;
proc reg data=bodyfat;
   model fat=skinfold thigh midarm;
```

```
.                              Sum of
Source              DF        Squares     F Value     Pr > F
Model                3      396.98461       21.52     <.0001
Error               16       98.40489
Corrected Total     19      495.38950
Root MSE                  2.47998    R-Square      0.8014
```

Something is useful.

```
.                            Parameter         Standard
Variable        DF            Estimate            Error
Intercept        1           117.08469         99.78240
skinfold         1             4.33409          3.01551
thigh            1            -2.85685          2.58202
midarm           1            -2.18606          1.59550

    t Value    Pr > |t|
       1.17      0.2578
       1.44      0.1699
      -1.11      0.2849
      -1.37      0.1896
```

But what? None of the p-values are significant.

Dilemma

The p-value for $F_{3,16}$ is 7×10^{-6}, 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. We have a seeming contradiction. What is the reason?

Look at the Extra SS. Here's how to do it in SAS.

```
proc reg data=bodyfat;
   model fat=skinfold thigh midarm /ss1 ss2;
```

```
Variable     Pr > |t|      Type I SS     Type II SS
skinfold       0.1699      352.26980       12.70489
thigh          0.2849       33.16891        7.52928
midarm         0.1896       11.54590       11.54590
```

Compare these values to the model SS

```
Model                         396.98461
Error                          98.40489
Corrected Total               495.38950
```

Interpretation

Type I:

- $SSM(skinfold) = 352.26$
- $SSM(thigh|skinfold) = 33.16$
- $SSM(midarm|skinfold, thigh) = 11.54$

Type II:

- $SSM(skinfold|thigh, midarm) = 12.70$
- $SSM(thigh|skinfold, midarm) = 7.52$

$Skinfold$ accounts for a lot of the SS by itself, but if $thigh$ and $midarm$ are included first, then $skinfold$ is redundant.

Notes

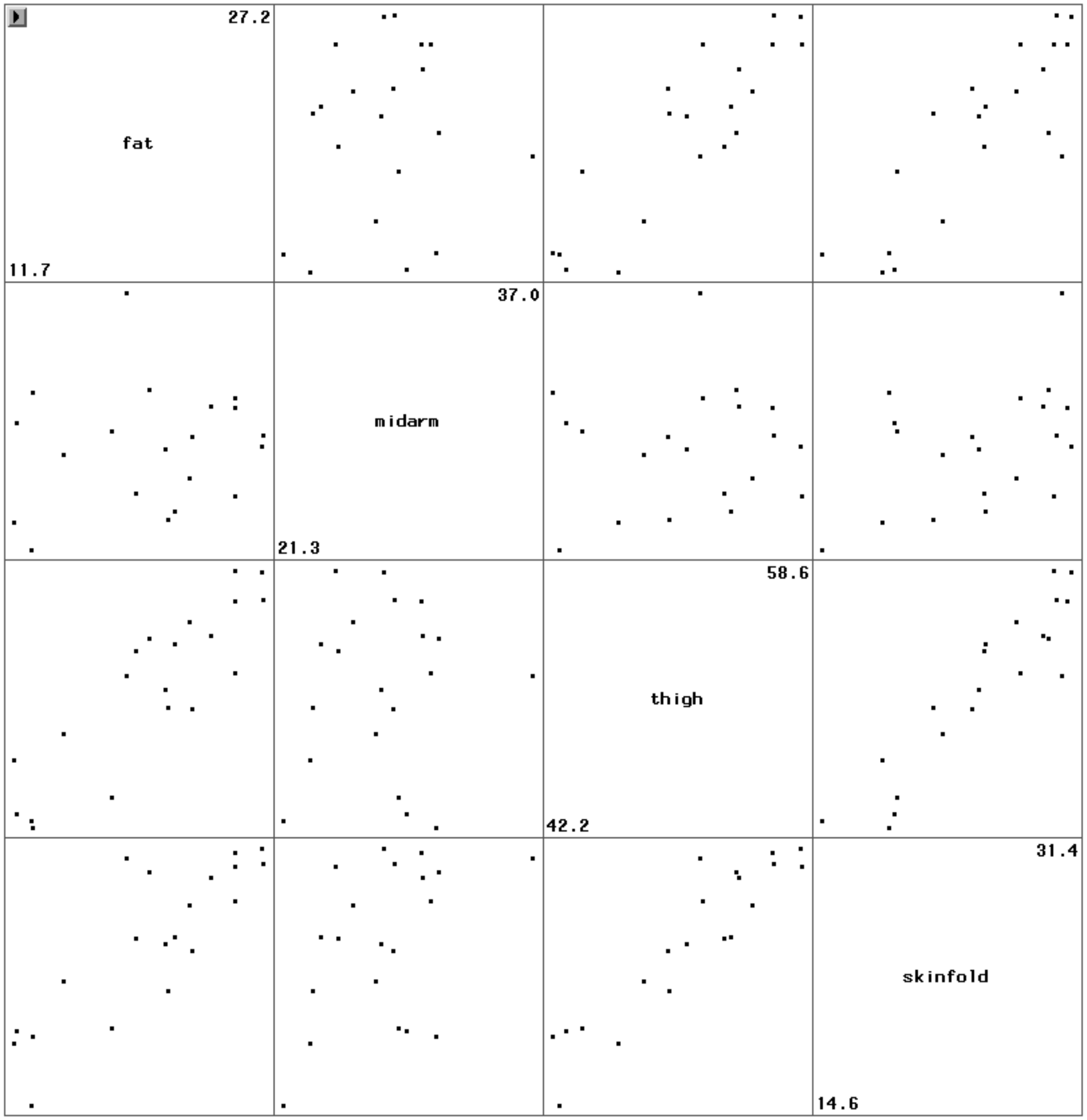
- Type I and Type II SS are very different for this example (except for the last one).
- The order in which we list the variables in the `model` statement affects the Type I SS, but does not affect the Type II SS.
- If we reorder the variables in the `model` statement we will get different Type I SS but the same Type II SS.

- Could variables be explaining the same SS and “canceling each other out”? We suspect so because the Type I and II SS are so different.
- We also observe parameter estimates that do not “make sense”: the regression coefficients for $thigh$ and $midarm$ are negative. (Large thighs = low body fat?)
- Run additional models to clarify the situation (more on this later).
- Also look at pairwise relationships, keeping in mind that more complicated relationships may exist.

```
proc corr data=bodyfat noprob;
```

```
            Pearson Correlation Coefficients, N = 20
               skinfold          thigh         midarm          fat
skinfold        1.00000        0.92384        0.45778      0.84327
thigh           0.92384        1.00000        0.08467      0.87809
midarm          0.45778        0.08467        1.00000      0.14244
fat             0.84327        0.87809        0.14244      1.00000
```

The correlation between $skinfold$ and $thigh$ is extremely high, so we most likely do not want to have both of these in our model. Both of these are also highly correlated with fat, which means they will be good predictors.


27.2
fat
11.7
37.0
midarm
21.3
58.6
thigh
42.2
31.4
skinfold
14.6

Rerun with single explanatory variables

```
proc reg data=bodyfat;
   model fat = skinfold;
   model fat = thigh;
   model fat = midarm;
```

```
.                    Parameter
Variable     DF       Estimate   t Value    Pr > |t|
Intercept     1       -1.49610     -0.45      0.6576
skinfold      1        0.85719      6.66      <.0001
Root MSE              2.81977    R-Square    0.7111

Intercept     1      -23.63449     -4.18      0.0006
thigh         1        0.85655      7.79      <.0001
Root MSE              2.51024    R-Square    0.7710

Intercept     1       14.68678      1.61      0.1238
midarm        1        0.19943      0.61      0.5491
Root MSE              5.19261    R-Square    0.0203
```

`test` statement in `proc reg`

This does any general linear test you specify. Note that the labels before the “`:`” are just names picked to identify the tests; they are useful, since you can do more than one test here and need to distinguish the output.

For example, the command

```
test thigh, midarm;
```

tests $\mathrm{H}_0 : \beta_2 = \beta_3 = 0$.

```
proc reg data=bodyfat;
   model fat=skinfold thigh midarm;
   skinonly: test thigh, midarm;
   thighonly:  test skinfold, midarm;
   skinmid:  test thigh;
```

```
   Test skinonly Results for Dependent Variable fat
                                 Mean
Source              DF         Square    F Value    Pr > F
Numerator            2       22.35741       3.64    0.0500
Denominator         16        6.15031
```

This is really borderline – probably should not remove both $thigh$ and $midarm$.

```
    Test thighonly Results for Dependent Variable fat
                                 Mean
Source             DF          Square    F Value    Pr > F
Numerator           2         7.50940       1.22    0.3210
Denominator        16         6.15031
```

This says a model with $thigh$ only would probably be okay.

```
.    Test skinmid Results for Dependent Variable fat
                                 Mean
Source             DF          Square    F Value    Pr > F
Numerator           1         7.52928       1.22    0.2849
Denominator        16         6.15031
```

This says a model with $skinfold$ and $midarm$ only, but not $thigh$, would probably be okay.

- How to pick the very best model depends on your objectives.
  - If your goal is only “black-box” prediction, with no interpretation on the coefficients, you might want to include some variables that are only marginally helpful.
  - But if your goal is to actually interpret the coefficient, you will also want to look at whether the coefficients make sense.
- For example, in the full model, the parameter estimates for $thigh$ and $midarm$ are both negative. This says, the larger your thighs and arms, the smaller your body fat, which doesn’t make much sense. This happens because $skinfold$ is already in the model, and so the positive influence of $thigh$ on fat has already been included.

Other uses

The `test` statement can be used to perform a significance test for any hypothesis involving a linear combination of the regression coefficients.

- Examples:

  $H_0 : \beta_4 = \beta_5$ (`test X4 = X5`)

  $H_0 : \beta_4 - 3\beta_5 = 12$ (`test X4-3*X5=12`)

- Separate equations by commas; if no $=$ sign appears then "`= 0`" is assumed.

Partial Correlations (r) – Coefficients of Partial Determination(r^2)

- Partial correlation measures the strength of a linear relation between two variables taking into account other variables.
- Recall from SLR that $r^2 = SSM/SST$. In contrast, now we are measuring the marginal increase in SS explained by including a new variable in the model, compared to the SSE.
- See page 270 for formulas.
- This is closely related to extra sums of squares.

Possible procedure to find partial correlation X_i, Y

This is not how you will actually calculate it, but it may help you visualize what is going on

- Predict Y using other X's
- Predict X_i using other X's
- Find correlation between the two sets of residuals
- Plotting the two sets of residuals vs each other on a scatterplot gives a partial correlation plot (later).

KNNL use the term *coefficient of partial determination* for the squared partial correlation r^2.

Partial correlations in SAS

```
proc reg data=bodyfat;
   model fat=skinfold thigh midarm / pcorr1 pcorr2;
```

```
                       Parameter Estimates
                   Squared Partial  Squared Partial
Variable     Pr > |t|    Corr Type I     Corr Type II
Intercept     0.2578               .                .
skinfold      0.1699         0.71110          0.11435
thigh         0.2849         0.23176          0.07108
midarm        0.1896         0.10501          0.10501
```

These use the Type I and Type II SS for computation:

- Type I Squared Partial Correlation uses Type I SS:
  $r^2 = \frac{SS1}{SS1+SSE}$
- Type II Squared Partial Correlation uses Type II SS:
  $r^2 = \frac{SS2}{SS2+SSE}$

Where e.g. for $thigh$,
$SS2 = SSM(thigh|skinfold, midarm)$, and
$SSE = SSE(skinfold, thigh, midarm)$.
We can similarly define partial correlations as the square root.
These give roughly the same information as Extra SS but are scaled.

Multicollinearity

Back to body fat example (KNNL page 256)

- The p-value for ANOVA F-test is $< .0001$.
- The p-values for the individual regression coefficients are 0.1699, 0.2849, and 0.1896.

None of these are near our standard significance level of 0.05. What is the explanation?

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.

If we solve the statistical problem then the numerical problem will also be solved. Our general goal is to refine a model that has redundancy in the explanatory variables. And we need to do this whether or not $X'X$ can be inverted without difficulty.

Extreme cases can help us understand the problems caused by multicollinearity.

- **One extreme**: Assume columns in X matrix were uncorrelated ($r = 0$). In that case Type I and Type II SS will be the same. The contribution of each explanatory variable to the model is the same whether or not the other explanatory variables are in the model. There is no overlap in the variation components explained by each variable.

- **Other extreme**: Suppose a linear combination of the explanatory variables is a constant (maybe a variable was accidentally included twice in the data set such that $X_1 = X_2$). The Type II SS for the X's involved will be zero because when one is included the other is redundant (it explains NO additional variation over the other variables).

Back to CS example: We did not have a big multicollinearity problem. Let’s fake a linear combination to illustrate a point.

```
data cs;
   infile ’H:\System\Desktop\csdata.dat’;
   input id gpa hsm hss hse satm satv genderm1;
data cs; set cs;
   sat=satm + satv;
proc reg data=cs;
   model gpa=sat satm satv;
```

```
.                              Sum of          Mean
Source                DF      Squares        Square    F Value    Pr > F
Model                  2      8.58384       4.29192       7.48    0.0007
Error                221    126.87895       0.57411
Corrected Total      223    135.46279
```

Something is wrong; $df_M = 2$, but there are 3 X's.

What is going on?

NOTE:

- Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.
- The following parameters have been set to 0, since the variables are a linear combination of other variables as shown: satv = sat - satm

```
                        Parameter      Standard
Variable      DF         Estimate         Error    Pr > |t|
Intercept      1          1.28868       0.37604      0.0007
sat            B      -0.00002456    0.00061847      0.9684
satm           B          0.00231       0.00110      0.0365
satv           0                0             .       .
```

SAS tells us the design matrix is not of full rank, i.e. its rank is $< p$. This means one column is a linear combination of some other columns. This in turn means that the determinant of $\mathbf{X}'\mathbf{X}$ is 0, and thus $\mathbf{X}'\mathbf{X}$ is not invertible.

Extent of Multicollinearity

This 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 $\mathbf{X}'\mathbf{X}$ does not have an inverse. We are mostly concerned with less extreme cases where $\mathbf{X}'\mathbf{X}$ does have an inverse, though the determinant may be quite small.

Effects of Multicollinearity

What happens when there is multicollinearity in the X's?

- Regression coefficients are not well estimated and may be meaningless, and similarly for standard errors of these estimates. (see chart on page 286)
- Type I SS and Type II SS will differ.
- R^2 and predicted values are usually okay.

Pairwise Correlations

Pairwise correlations can be used to check for “pairwise” collinearity. This is useful but note that they do not show more complicated linear dependence.

$$\begin{aligned} Corr(skinfold, thigh) &= 0.9238 \\ Corr(skinfold, midarm) &= 0.4578 \\ Corr(thigh, midarm) &= 0.0847 \end{aligned}$$

But this is not the whole story. For example, if we think of $skinfold$ as a Y variable and model it with both $midarm$ and $thigh$ as explanatory variables, we get a very large $R^2 = 0.9986$, meaning that $skinfold$ is almost exactly predicted from the other two. Simple pairwise correlations do not pick up on this, but the regression results tell us something like this was going on.

Similarly in the example with sat the problem was with a set of three variables: sat, $satv$, and $satm$. Any two of these would be okay, but three is impossible. We would not have been aware of the extent of this problem just from the pairwise correlations.

```
proc corr data=cs noprob;
var sat satm satv;
```

```
                  sat            satm            satv
sat           1.00000         0.84450         0.86623
satm          0.84450         1.00000         0.46394
satv          0.86623         0.46394         1.00000
```

Polynomial Regression

- We can fit a quadratic, cubic, etc. relationship by defining squares, cubes, etc., in a data step and using them as additional explanatory variables.
- We can also do this with more than one explanatory variable, in which case we also often include an interaction term.
- When we do this we generally create a multicollinearity problem, which can often be corrected by standardization.

Warning: You know that with linear regression models it is a bad idea to extrapolate beyond the scope of the data. In polynomial regression it is a VERY bad idea.

KNNL Example, page 300 (`nknw302.sas`)

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: levels of charge and temperature are planned.
- The experimenter wants to know if a linear or quadratic function is appropriate, and if an interaction between charge rate and temperature should be included in the model.
  - An interaction means that the way in which charge rate affects the life of the cell depends on the temperature.
  - A model with no interaction is called *additive*.
  - MUCH more on this in ANOVA.

Input and check the data.

```
data powercell;
   infile 'H:\System\Desktop\CH07TA09.DAT';
   input cycles chrate temp;
proc print data=powercell;
```

```
Obs     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
```

Use a `data` step to create new variables and run the regression with polynomial and interaction terms.

```
data powercell; set powercell;
  chrate2=chrate*chrate;
  temp2=temp*temp;
  ct=chrate*temp;
proc reg data=powercell;
   model cycles=chrate temp chrate2 temp2 ct / ss1 ss2;
```

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

                Parameter
Variable   DF    Estimate  Pr > |t|   Type I SS   Type II SS
Intercept   1   337.72149    0.0741      325424   5315.62944
chrate      1  -539.51754    0.1011       18704   4220.41673
temp        1     8.91711    0.3761       34202    988.38036
chrate2     1   171.21711    0.2359  1645.96667   1901.19474
temp2       1    -0.10605    0.6244   284.92807    284.92807
ct          1     2.87500    0.5092   529.00000    529.00000
```

Conclusion

Overall F is significant, but no individual t's are significant $\rightarrow$ multicollinearity problem. Also notice the extreme difference between Type I and Type II SS for $chrate$ and $temp$.
Look at the correlations (`proc corr`) for a clue :

```
        Pearson Correlation Coefficients, N = 11

             chrate       temp    chrate2      temp2         ct
chrate      1.00000    0.00000    0.99103    0.00000    0.60532
temp        0.00000    1.00000    0.00000    0.98609    0.75665
chrate2     0.99103    0.00000    1.00000    0.00592    0.59989
temp2       0.00000    0.98609    0.00592    1.00000    0.74613
ct          0.60532    0.75665    0.59989    0.74613    1.00000
```

There are some very high correlations

$$\begin{aligned} r(chrate, chrate2) &= 0.99103 \\ r(temp, temp2) &= 0.98609 \end{aligned}$$

Correlation between powers of variables are causing a big multicollinearity problem.

A remedy

- We can remove or reduce the correlation between explanatory variables and their powers by *centering*.
- Centering means that you subtract off the mean from each variable.
- We often rescale by standardizing (subtract the mean and divide by the standard deviation), but subtracting the mean is the key.

Use `proc standard` to center the explanatory variables. (`proc standard` will standardize a dataset to any given mean and std dev; we set a mean of 0 to center the variables. To set a given standard deviation, we could also say `std=1`, for example. The same names are used for the standardized variables.)

Recompute the squares, cubes, etc., using the centered variables, and rerun the regression analysis.

Note the use of “`drop`” in the `data` step. (opposite is “`keep`”)

```
data copy; set powercell;
   schrate=chrate; stemp=temp;
   drop chrate2 temp2 ct;
*center the variables and put them in dataset std;
proc standard data=copy out=std mean=0;
   var schrate stemp;
* schrate and stemp now have mean 0;
proc print data=std;
```

```
Obs    cycles    chrate    temp    schrate    stemp
  1      150       0.6      10      -0.4       -10
  2       86       1.0      10       0.0       -10
  3       49       1.4      10       0.4       -10
  4      288       0.6      20      -0.4         0
  5      157       1.0      20       0.0         0
  6      131       1.0      20       0.0         0
  7      184       1.0      20       0.0         0
  8      109       1.4      20       0.4         0
  9      279       0.6      30      -0.4        10
 10      235       1.0      30       0.0        10
 11      224       1.4      30       0.4        10
```

Now both $schrate$ and $stemp$ have mean 0.

Recompute squares and interaction term using standardized variables.

```
data std; set std;
  schrate2=schrate*schrate;
  stemp2=stemp*stemp;
  sct=schrate*stemp;
```

```
.                   Parameter Estimates
                  Parameter      Standard
Variable           Estimate         Error   Pr > |t|
Intercept         151.42544      45.45653     0.0208
chrate           -139.58333      33.04176     0.0083
temp                7.55000       1.32167     0.0023
schrate2          171.21711     127.12550     0.2359
stemp2             -0.10605       0.20340     0.6244
sct                 2.87500       4.04677     0.5092
     Type I SS    Type II SS
        325424         11631
         18704         18704
         34202         34202
    1645.96667    1901.19474
     284.92807     284.92807
     529.00000     529.00000
```

Now we can see that the linear terms are significant and the quadratic terms are not. See book: general linear test shows that quadratic and interaction terms can be omitted. Things are much more clear now and we can trust our results since the correlations among X's are virtually eliminated.

Type I and Type II SS are almost the same.

```
proc corr data=std noprob;
var chrate temp schrate2 stemp2 sct;
```

```
        Pearson Correlation Coefficients, N = 11
              chrate         temp     schrate2       stemp2
chrate       1.00000      0.00000      0.00000      0.00000
temp         0.00000      1.00000      0.00000      0.00000
schrate2     0.00000      0.00000      1.00000      0.26667
stemp2       0.00000      0.00000      0.26667      1.00000
sct          0.00000      0.00000      0.00000      0.00000
```

In this particular example, the correlations went to zero after centering. That happened because all possible combinations were equally represented (in ANOVA terms, a balanced design). Correlations will not always go to zero after centering, but they should be reduced.

A typical quadratic regression model with two X variables would be

$$\begin{aligned} Y_i = & \quad \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_{1,1}(X_{i,1} - \bar{X}_1)^2 + \\ & \beta_{2,2}(X_{i,2} - \bar{X}_2)^2 + \beta_{1,2}(X_{i,1} - \bar{X}_1)(X_{i,2} - \bar{X}_2) + \epsilon_i \end{aligned}$$

Notice the alternate convention for subscripts sometimes used on the regression coefficients.

Comment on General Linear Test and Extra SS

The test for all second-order terms $= 0$ uses $SSM(schrate2, stemp2, sct|chrate, temp)$. This is not given directly from our SAS output. We can use the usual `test` statement in `proc reg`:

```
second:  test schrate2, stemp2, sct;
```

```
  Test second Results for Dependent Variable cycles
                               Mean
Source          DF           Square    F Value    Pr > F
Numerator        3        819.96491       0.78    0.5527
Denominator      5       1048.08772
```

But we can also get the necessary SS from $SS1$ with a little work:

$$SSM(schrate2, stemp2, sct|schrate, stemp) = \\ SSM(full) - SSM(schrate, stemp)$$

Notice that the Type I SS are

$$\begin{aligned}
SS1(schrate2) &= SSM(schrate2|chrate, temp) \\
&= SSM(chrate, temp, schrate2) - \\
&\quad SSM(chrate, temp) \\
SS1(stemp2) &= SSM(stemp2|chrate, temp, schrate2) \\
&= SSM(chrate, temp, schrate2, stemp2) - \\
&\quad SSM(chrate, temp, schrate2) \\
SS1(sct) &= SSM(sct, chrate, temp, schrate2, stemp2) \\
&= SSM(chrate, temp, schrate2, stemp2, sct) - \\
&\quad SSM(chrate, temp, schrate2, stemp2)
\end{aligned}$$

$$\begin{aligned}
&SS1(schrate2) + SS1(stemp2) + SS1(sct) \\
&\quad = \quad SSM(chrate, temp, schrate2, stemp2, sct) \\
&\qquad\quad -SSM(chrate, temp) \\
&\quad = \quad SSM(schrate2, stemp2, sct | chrate, temp)
\end{aligned}$$

Thus, we can get the required extra SS from adding certain Type I SS. This only works if the second-order terms (in general, the terms to be tested) are listed last in the `model` statement.

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 (0/1) and one continuous variable (Section 8.3 - doing this out of sequence)
- Two continuous variables

First Special Case: One binary variable and one continuous variable

(From Section 8.3)

X_1 takes values 0 and 1 corresponding to two different groups or categories.

X_2 is a continuous variable.

Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$

As you will see, this is a convenient way of writing down two separate SLR models for the two categories.

When $X_1 = 0$, the model becomes

$$\begin{aligned} Y &= \beta_0 + \beta_1(0) + \beta_2 X_2 + \beta_3(0) X_2 + \epsilon \\ &= \beta_0 + \beta_2 X_2 + \epsilon \end{aligned}$$

This is a SLR model for Y as a function of X_2 with intercept β_0 and slope β_2.

When $X_1 = 1$, the model becomes

$$\begin{aligned} Y &= \beta_0 + \beta_1(1) + \beta_2 X_2 + \beta_3(1)X_2 + \epsilon \\ &= \beta_0 + \beta_1 + \beta_2 X_2 + \beta_3 X_2 + \epsilon \\ &= (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X_2 + \epsilon \end{aligned}$$

This is a SLR model for Y as a function of X_2 with intercept $(\beta_0 + \beta_1)$ and slope $(\beta_2 + \beta_3)$.

These are both SLR models, and we could model them separately (make one data set for each group). But we can more easily model the combination of the two with a multiple regression model using a *binary (0-1) variable*, and this allows us to look at both groups together. It has the added benefit of using all the data together to get the variance estimate, increasing our error degrees of freedom.

Some useful tests in this situation include

- $H_0 : \beta_1 = \beta_3 = 0$ is the hypothesis that the regression lines are the same.
- $H_0 : \beta_1 = 0$ is the hypothesis that the two intercepts are equal.
- $H_0 : \beta_3 = 0$ is the hypothesis that the two slopes are equal.

KNNL Example, page 315

(`nknw459.sas`)

- Y is number of months it takes for an insurance company to adopt an innovation.
- X_1 is the type of firm (a qualitative or categorical variable): X_1 takes the value 0 if it is a mutual fund firm and 1 if it is a stock fund firm.
- X_2 is the size of the firm (a continuous variable).

We ask whether stock firms adopt the innovation slower or faster than mutual firms. We ask the question across all firms, regardless of size.

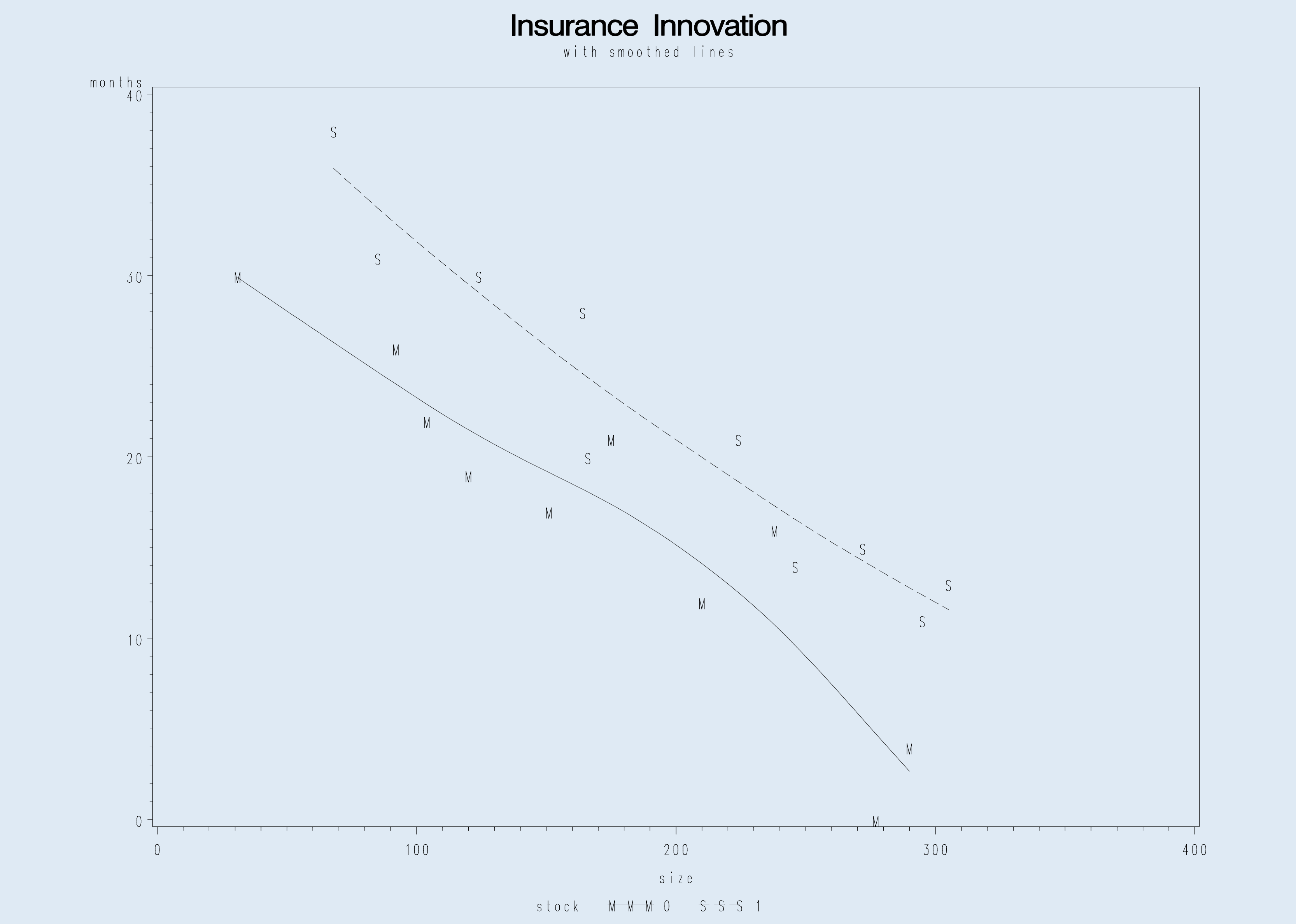
```
data insurance;
   infile 'H:\System\Desktop\Ch11ta01.dat';
   input months size stock;
proc print data=insurance;
```

```
Obs     months     size     stock
  1       17        151        0
  2       26         92        0
...
 19       30        124        1
 20       14        246        1
```

Plot the data with two symbols:

```
symbol1 v=M i=sm70 c=black l=1;
symbol2 v=S i=sm70 c=black l=3;
title1 'Insurance Innovation';
proc sort data=insurance;
    by stock size;
title2 'with smoothed lines';
proc gplot data=insurance;
    plot months*size=stock;
```

Note the use of a new syntax in the `plot` statement to plot separate lines for different values of the categorical variable stock.


Insurance Innovation
with smoothed lines
months
40
30
20
10
0
0
100
200
300
400
size
stock M M M 0 S S S 1

Interaction Effects

- Interaction expresses the idea that the effect of one explanatory variable on the response depends on another explanatory variable.
- In this example, this would mean that the slope of the line depends on the type of firm.
- Are both lines the same?
- Use the `test` statement to test whether $\beta_1 = \beta_3 = 0$.
- Make the X_3 variable which is the product of stock (X_1) and size (X_2).

```
data insurance; set insurance;
   sizestock=size*stock;

proc reg data=insurance;
   model months = stock size sizestock;
   sameline: test stock, sizestock;
```

```
                        Sum of               Mean
Source                 Squares             Square     Pr > F
Model               1504.41904          501.47301     <.0001
Error                176.38096           11.02381
Corrected Total     1680.80000
Root MSE               3.32021     R-Square        0.8951
```

```
.   Test sameline Results for Dependent Variable months
                                    Mean
Source              DF            Square     F Value     Pr > F
Numerator            2         158.12584       14.34     0.0003
Denominator         16          11.02381
```

The lines are not the same. How are they different?

```
.                Parameter          Standard
Variable          Estimate             Error    Pr > |t|
Intercept         33.83837           2.44065      <.0001
stock              8.13125           3.65405      0.0408
size              -0.10153           0.01305      <.0001
sizestock     -0.00041714            0.01833      0.9821
```

It appears we can ignore β_3 : Two parallel lines.

```
proc reg data=insurance;
   model months = stock size;
```

```
                Analysis of Variance
                                   Sum of
Source                  DF        Squares     Pr > F
Model                    2     1504.41333     <.0001
Error                   17      176.38667
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
```

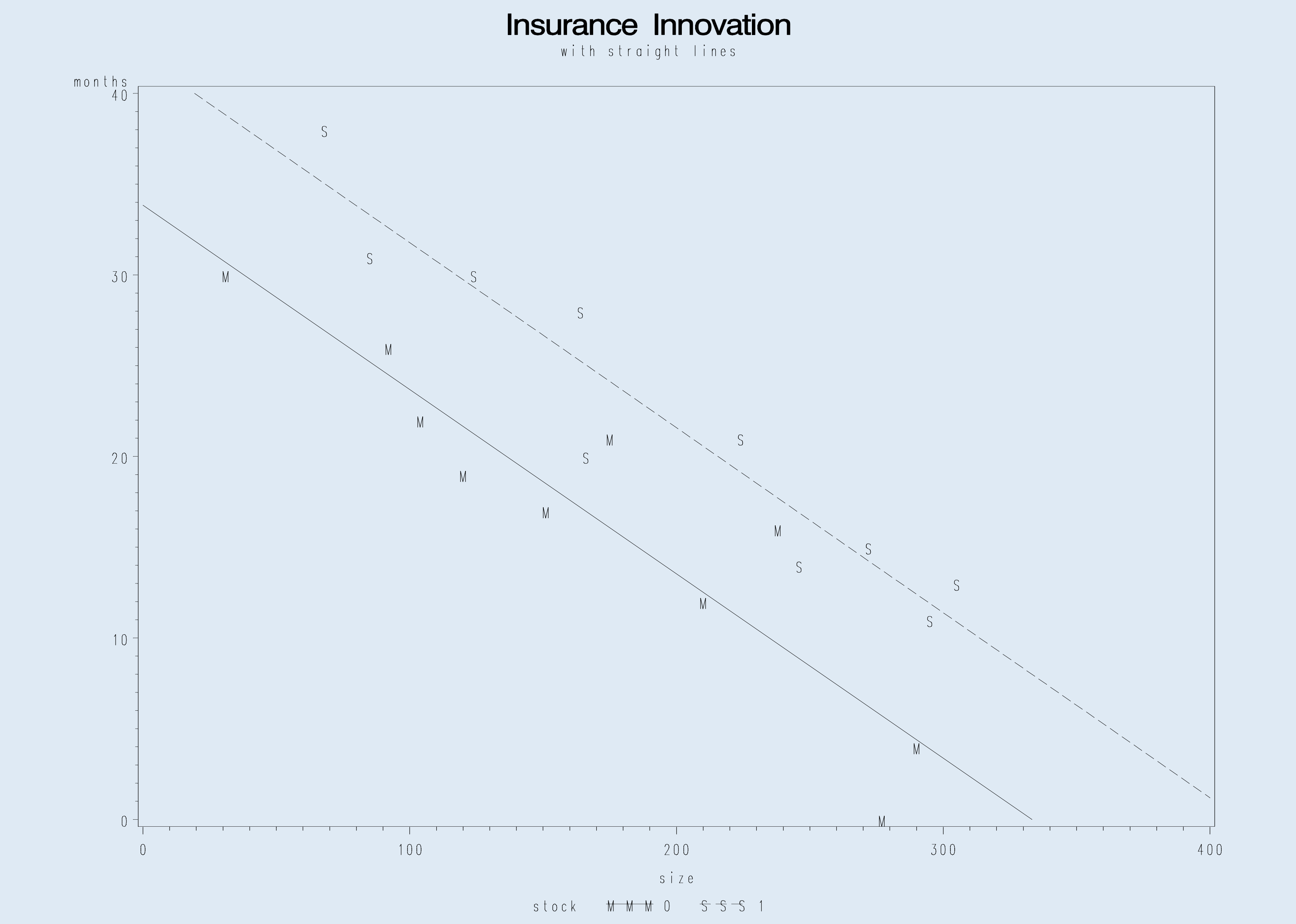
```
.                       Parameter Estimates
               Parameter      Standard
Variable        Estimate          Error    Pr > |t|
Intercept       33.87407        1.81386      <.0001
size            -0.10174        0.00889      <.0001
stock            8.05547        1.45911      <.0001
```

- The two lines have different intercepts, but it is safe to assume they have the same slope.
- For mutual fund firms, $X_1 = 0$,
  - the model is $Y = \beta_0 + \beta_2 X_2 + \epsilon$;
  - the fitted equation is $\hat{Y} = 33.784 - 0.102 X_2$.
- For stock firms, $X_1 = 1$,
  - the model is $Y = \beta_0 + \beta_1 + \beta_2 X_2 + \epsilon$;
  - the estimated intercept is $(33.874 + 8.055) = 41.93$, so the fitted equation is $\hat{Y} = 41.93 - 0.102 X_2$.

Plot with two regression lines (note this manner of plotting will permit different slopes; to get the actual lines given by the model we would have to overlay the predicted values on the plot, as we did in Homework 3).

```
symbol1 v=M i=rl c=black;
symbol2 v=S i=rl c=black;

proc gplot data=insurance;
   plot months*size=stock;
```


Insurance Innovation
with straight lines
months
40
30
20
10
0
0
100
200
300
400
size
stock M M M 0 S S S 1

Second Special Case: Two Continuous Variables

The model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$ can be rewritten as follows:

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ Y &= \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2 + \epsilon \end{aligned}$$

The coefficient of one explanatory variable depends on the value of the other explanatory variable.

Variable Selection and Model Building

We usually want to choose a model that includes a subset of the available explanatory variables.

Two separate but related questions:

- How many explanatory variables should we use (i.e., subset size)? Smaller sets are more convenient, but larger sets may explain more of the variation (SS) in the response.
- Given the subset size, which variables should we choose?

Criteria for Model Selection

To determine an appropriate subset of the predictor variables, there are several different criteria available.

- We will go through them, noting their benefits and drawbacks.
- They include R^2, adjusted R^2, Mallow's C_p, MSE, $PRESS$, AIC, SBC.
- SAS will provide these statistics, so you should pay more attention to what they are good for than how they are computed.
- To obtain them from SAS, place after the `model` statement `/selection = MAXR ADJRSQ CP`.
- Note that the different criterion may not lead to the same model in every case.

R^2 and Adjusted R^2 (or MSE) Criterion

- The text uses $R_p^2 = R^2 = 1 - \frac{SSE}{SSTO}$ (see page 354). Their subscript is just the number of variables in the associated model.
- The goal in model selection is to maximize this criterion.
    - One MAJOR drawback to R^2 is that the addition of any variable to the model (significant or not) will increase R^2 (perhaps not enough to notice depending on the variable). At some point, added variables just get in the way!

- The *Adjusted R^2 criterion* penalizes the R^2 value based on the number of variables in the model. Hence it eventually starts decreasing as unnecessary variables are added.

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}$$

(we end up subtracting off more as p is increased)

- Maximizing the Adjusted R^2 criterion is one way to select a model. As the text points out this is equivalent to minimizing the MSE since

$$\begin{aligned} R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO} &= 1 - \frac{MSE}{SSTO/(n-1)} \\ &= 1 - \frac{MSE}{\text{constant}} \end{aligned}$$

Mallow's C_p Criterion

- The basic idea is to compare subset models with the full model.
- The full model is good at prediction, but if there is multicollinearity our interpretations of the parameter estimates may not makes sense. A subset model is good if there is not substantial "bias" in the predicted values (relative to the full model).
- The C_p criterion looks at the ratio of error SS for the model with p variables to the MSE of the full model, then adds a penalty for the number of variables.

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

- SSE is based on a specific choice of $p-1$ variables (p is the number of regression coefficients including the intercept); while MSE is based on the full set of variables.
- A model is good according to this criterion if $C_p \leq p$. We may choose the smallest model for which $C_p \leq p$, so a benefit of this criterion is that it can achieve for us a "good" model containing as few variables as possible.
- One might also choose to pick the model that minimizes C_p.
- See page 357-359 for details.

PRESS Statistic

- Stands for PREdiction Sums of Squares
- Obtained by the following algorithm: For each observation i, delete the observation and predict Y for that observation using a model based on the $n-1$ cases. Then look at SS for the observed minus predicted.

$$PRESS_p = \sum (Y_i - \hat{Y}_{i(i)})^2$$

- Models with small PRESS statistic are considered good candidates.

SBC and AIC

Criterion based on $-2 * log(likelihood)$ plus a penalty for more complexity.

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

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

Note that different criteria will not give the identical answer.

Model Selection Methods

There are three commonly available. To apply them using SAS, you use the option `selection = *****` after the `model` statement. The methods include

- *Forward Selection* (`FORWARD`) - starts with the null model and adds variables one at a time.
- *Backward Elimination* (`BACKWARD`) - starts with the full model and deletes variables one at a time.
- *Forward Stepwise Regression* (`STEPWISE`) - starts with the null model and checks for adds/deletes at each step. This is probably the preferred method (see section on multicollinearity!). It is forward selection, but with a backward glance at each step.

These methods all add/delete variables based on Partial F-tests. (See page 364)

Some additional options in the `model` statement

- `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
- `START=n` limits output to models that include at least n explanatory variables

Ordering Models of the Same Subset Size

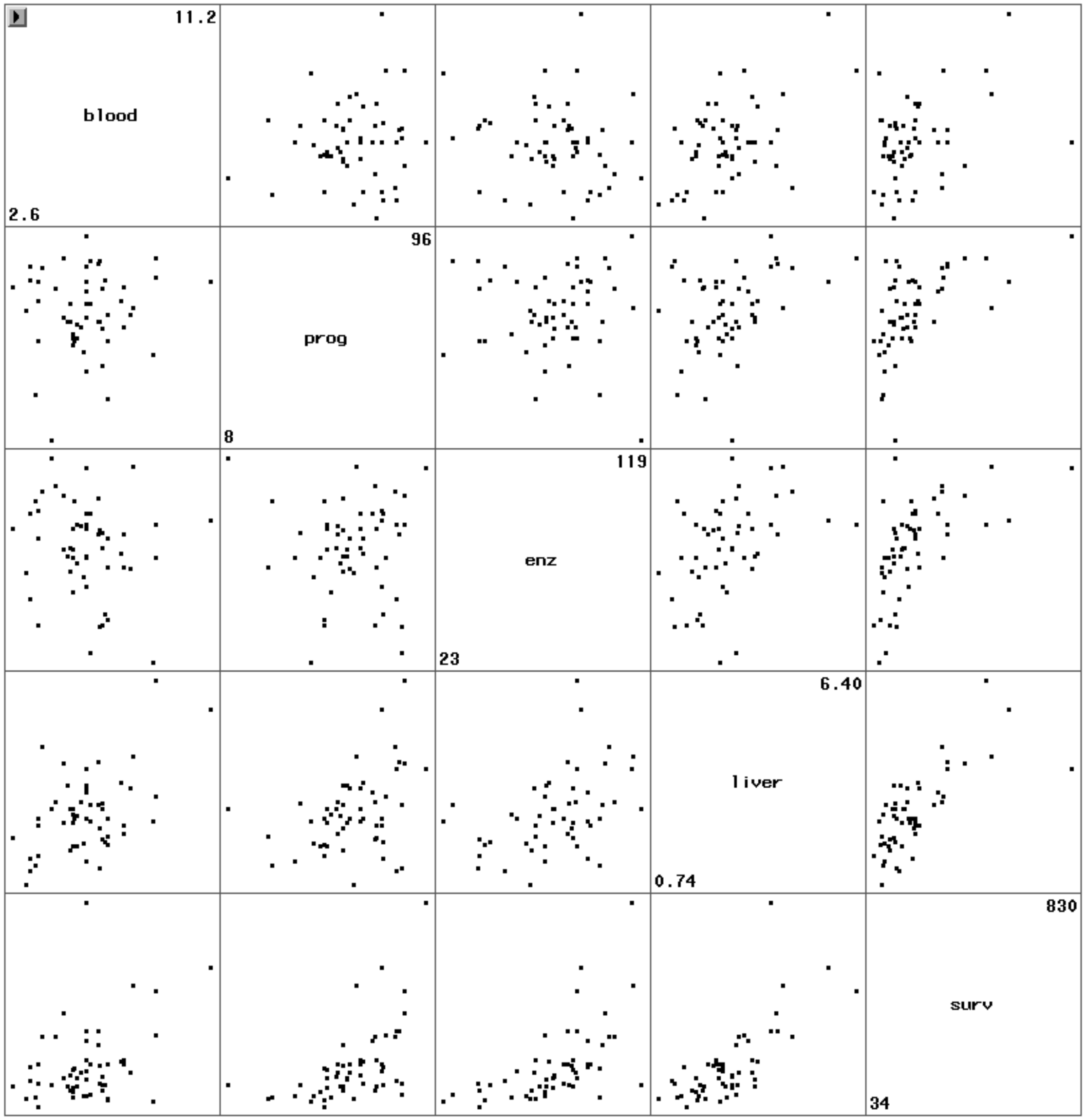
- Use R^2 or SSE.
- This approach can lead us to consider several models that give us approximately the same predicted values.
- May need to apply knowledge of the subject matter to make a final selection.
- If prediction is the key goal, then the choice of variables is not as important as if interpretation is the key.

Surgical Unit Example

- References: KNNL Section 9.2 (p350ff), `nknw334.sas`
- Y is the survival time
- Potential X's include Blood clotting score (X_1), Prognostic Index (X_2), Enzyme Function Test (X_3), and Liver Function Test (X_4).
- $n = 54$ patients were observed.

Initial diagnostics note curved lines and non-constant variance, suggesting that Y should be transformed with a $\log$. Take a look at the plots in the SAS file and play with some analyses on your own.

```
data surgical;
   infile 'H:\System\Desktop\Ch08ta01.dat';
   input blood prog enz liver surv;
```



blood
11.2
2.6
prog
96
8
enz
119
23
liver
6.40
0.74
surv
830
34

Take the $\log$ of survival

```
data surgical;
   set surgical;
   lsurv=log(surv);
proc reg data=surgical;
   model lsurv=blood prog enz liver/
   selection=rsquare cp aic sbc b best=3;
```

```
Number in
  Model     R-Square         C(p)          AIC          SBC
      1       0.5274     787.9471     -87.3085    -83.33048
      1       0.4424     938.6707     -78.3765    -74.39854
      1       0.3515     1099.691     -70.2286    -66.25061
-------------------------------------------------------------
      2       0.8129     283.6276    -135.3633   -129.39638
      2       0.6865     507.8069    -107.4773   -101.51034
      2       0.6496     573.2766    -101.4641    -95.49714
-------------------------------------------------------------
      3       0.9723       3.0390    -236.5787   -228.62281
      3       0.8829     161.6520    -158.6434   -150.68745
      3       0.7192     451.8957    -111.4189   -103.46299
-------------------------------------------------------------
      4       0.9724       5.0000    -234.6217   -224.67680
```

- One model stands out: the first one with 3 variables ($C_p = 3.04 < p = 4$).
- The full model has $C_p = 5 = p$.
- The parameter estimates indicate that the desired model is the one with $blood$, $prog$ and enz, but not $liver$.

```
Number             -----------------Parameter Estimates-----------------
Model    R-Square   Intercept     blood       prog       enz       liver
     1     0.5274     3.90609         .         .         .      0.42771
     1     0.4424     3.55863         .         .    0.01973          .
     1     0.3515     3.68138         .    0.02211         .          .
------------------------------------------------------------------------
     2     0.8129     2.08947d        .    0.02271    0.02015         .
     2     0.6865     3.19784         .         .     0.01301    0.32010
     2     0.6496     3.24325         .    0.01403         .     0.34596
------------------------------------------------------------------------
     3     0.9723     1.11358   0.15940    0.02140    0.02193         .
     3     0.8829     2.16970         .    0.01819    0.01612    0.18846
     3     0.7192     2.68966   0.09239          .    0.01604    0.22556
------------------------------------------------------------------------
     4     0.9724     1.12536   0.15779    0.02131    0.02182    0.00442
```

In this particular example you would probably come to the same conclusion based on the Type II SS, but not on the individual correlations: $liver$ has the *highest* individual correlation with $lsurv$ (but also is correlated with the other three).

Below we see that this is the same model chosen by forward stepwise regression:

```
proc reg data=surgical;
   model lsurv=blood prog enz liver / selection=stepwise;
```

```
.           Summary of Stepwise Selection
        Variable     Variable     Number     Partial
Step    Entered      Removed      Vars In    R-Square
  1     liver                        1        0.5274
  2     enz                          2        0.1591
  3     prog                         3        0.1964
  4     blood                        4        0.0895
  5                  liver           3        0.0000

Model
R-Square      C(p)      F Value    Pr > F
 0.5274     787.947      58.02     <.0001
 0.6865     507.807      25.89     <.0001
 0.8829     161.652      83.83     <.0001
 0.9724      5.0000     158.65     <.0001
 0.9723      3.0390       0.04     0.8442
```