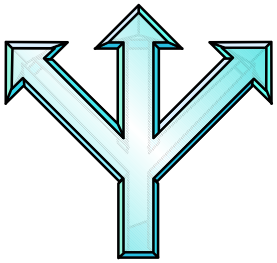
Statistical Techniques II

EXST7015

Multiple Regression Variable Selection



Variable Selection

- We have previously discussed the concept of partial sums of squares and partial regression coefficients.
- As you know, the addition or removal of any variable will change all other variables in the model.
- Therefore, if you decide to add or remove variables from a model this should be done one variable at a time.

Stepwise Variable Selection

- The procedure has been formalized in several options. We will discuss a few of these, Forward Selection, Backward Selection and Stepwise Selection.
- One additional reason for reducing the model in Example 2 is that we had multicollinearity. Stepwise regression is not specifically designed to avoid multicollinearity, but it will tend to not pick up two variables that are collinear.

Stepwise Selection *(continued)*

- **Backward selection is the simplest.**
  - **It starts with the full model, a model with all variables of interest already present in the model.**
  - **A selection criteria is established. Perhaps we want no non-significant variables in the model (α=0.05).**

Stepwise Selection *(continued)*

- **Backward selection (continued)**
  - **Step 1: The least significant variable is examined with a F test of the Type II SS (or t-test of the regression coefficient).**
    - **If the significance level does not meet our criteria, the variable is deleted from the model. The model is then refit.**
  - **Step 2: If one of the remaining variables does not meet the selection criteria it is removed. The model is refit without the variable.**

Stepwise Selection *(continued)*

- **Backward selection (continued)**
  - **These steps are repeated until all variables in the model meet the selection criteria.**
- **SAS options on the model statement**
  - **Choose backward selection with the "selection=backward" option.**
  - **SLSTAY = 0.# , chooses a significance level to keep in the model. The default value is 0.10.**

Stepwise Selection *(continued)*

- **Backward selection (continued)**
  - **See the SAS output. For our example the first variable removed was CENSUS. Subsequently the analysis removed AGE, NOBEDS and NURSES.**
  - **The SAS summary reports the outcome of the backward selection.**

```
►All variables left in the model are significant at the 0.1000 level.
►
►Summary of Backward Elimination Procedure for Dependent Variable INFRISK
►        Variable   Number   Partial     Model
►Step    Removed        In      R**2      R**2         C(p)           F   Prob>F
►   1    CENSUS          7    0.0002    0.5249       7.0440      0.0440   0.8343
►   2    AGE             6    0.0023    0.5226       5.5431      0.5037   0.4795
►   3    NOBEDS          5    0.0029    0.5197       4.1729      0.6386   0.4260
►   4    NURSES          4    0.0036    0.5161       2.9592      0.7999   0.3731
```

Stepwise Selection *(continued)*

- **Backward selection (continued)**
  - **The final model had 4 variables and the intercept (SAS will not remove the intercept).**

```
►
►Step 4   Variable NURSES Removed    R-square = 0.51613081   C(p) =  2.95916112
►
►                DF         Sum of Squares      Mean Square          F   Prob>F
►Regression       4           103.93833183      25.98458296      28.80   0.0001
►Error          108            97.44149118       0.90223603
►Total          112           201.37982301
►
►                Parameter        Standard          Type II
►Variable         Estimate           Error   Sum of Squares          F   Prob>F
►INTERCEP      -0.06358059      0.53320703       0.01282855       0.01   0.9053
►LTOFSTAY       0.18841053      0.05471423      10.69867196      11.86   0.0008
►CULRATIO       0.04644573      0.00992331      19.76510132      21.91   0.0001
►XRAY           0.01205242      0.00535081       4.57750739       5.07   0.0263
►SERVICES       0.02046537      0.00634744       9.37911740      10.40   0.0017
```

Stepwise Selection *(continued)*

- **Backward selection (continued)**
  - **All are significant at the 0.10 level of α, the requested (default) criteria (except for the intercept) .**

```
►
►Step 4   Variable NURSES Removed    R-square = 0.51613081   C(p) =  2.95916112
►                DF         Sum of Squares      Mean Square          F   Prob>F
►Regression       4           103.93833183      25.98458296      28.80   0.0001
►Error          108            97.44149118       0.90223603
►Total          112           201.37982301
►
►                Parameter        Standard          Type II
►Variable         Estimate           Error   Sum of Squares          F   Prob>F
►INTERCEP      -0.06358059      0.53320703       0.01282855       0.01   0.9053
►LTOFSTAY       0.18841053      0.05471423      10.69867196      11.86   0.0008
►CULRATIO       0.04644573      0.00992331      19.76510132      21.91   0.0001
►XRAY           0.01205242      0.00535081       4.57750739       5.07   0.0263
►SERVICES       0.02046537      0.00634744       9.37911740      10.40   0.0017
```

Stepwise Selection *(continued)*

- **Forward stepwise selection works by calculating all possible simple linear regressions, and picking the best one to start with.**
  - **Again, the F test of the Type II SS, or t-test of the slopes, are used as criteria for selection.**
  - **The "best" variable is the most significant one, as long as it meets a minimum criteria.**
  - **Once chosen, this best variable will remain in the model for the whole analysis.**

Stepwise Selection *(continued)*

- **Forward stepwise selection (continued)**
  - **After picking the one best variable, the analysis checks all possible 2 factor models, trying each of the remaining variables together with the first one chosen.**
  - **If there are additional variables that meet the criteria, the analysis chooses the best of these.**
  - **The step is repeated until no remaining variables meet the criteria.**

Stepwise Selection *(continued)*

- SAS model statement options
  - Forward selection is requested with the "selection=forward" option
  - the minimum criteria can be set with the SLENTRY = 0.# option, which has a default value of 0.50.
  - Forward selection has one limitation. Once a variable is selected for inclusion in the model it will not be removed, even if the entry of a later variable causes it to become not significant.

Stepwise Selection *(continued)*

- **There is a variation of this called the "Stepwise" option requested by "selection=stepwise".**
  - **This is like forward stepwise selection, except that at each step the analysis checks to make sure that each variable still meets the criteria. If a variable falls below the criteria it will be removed.**
  - **Think of it a *forward selection with a backward glance*.**

Stepwise Selection *(continued)*

- This "stepwise" option can use both options previously mentioned, SLEntry = 0.15 (default), and SLStay = 0.15 (default).
- It is the analysis I included on the handout.

```
All variables left in the model are significant at the 0.1500 level.
No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Procedure for Dependent Variable INFRISK
       Variable        Number   Partial    Model
Step   Entered   Removed   In      R**2     R**2      C(p)          F   Prob>F
   1   CULRATIO             1    0.3127   0.3127   41.5161    50.4918   0.0001
   2   LTOFSTAY             2    0.1377   0.4504   13.3525    27.5690   0.0001
   3   SERVICES             3    0.0430   0.4934    5.9368     9.2513   0.0029
   4   XRAY                 4    0.0227   0.5161    2.9592     5.0735   0.0263
```

Stepwise Selection *(continued)*

- Two approaches were used (backward and stepwise) and both produced the same model. This is often, but not always, true. I usually use the "stepwise" option, but I like to check backward too.
- Notice that SAS has a column for variables entered and variables removed. No variables were removed in this particular analysis.

Stepwise Selection *(continued)*

- There is one additional option that can be useful among the selection options. You can specify KEEP=#. This will force SAS to keep the first # variables in the model. They will be in to start with and will not be removed.
- This is good if you have an base model you want to keep intact and want to check for additional variables.

Stepwise Selection *(continued)*

- You will notice that the summary table produced by SAS contains a number of diagnostics in addition to the F value and its probability. These include a Partial R^2, Adjusted R^2, and Mallow's $C_{(p)}$ statistic.
- These are additional tools that can be used in variable selection.
- You are already familiar with the Partial R^2, the amount of the remaining variation accounted for by a variable.

Stepwise Variable Selection *(continued)*

- **The others are used to determine the number of variables to keep in the model.**
- **As you know, when a variable is added to a model the usual R^2 always gets larger. The adjusted R^2 is "adjusted" such that the value will not get larger unless the variation accounted for by the variable is equal to at least one MSE. Otherwise this value can actually decrease.**
- **SAS no longer includes these values.**

Stepwise Selection *(continued)*

- I ran the following program to force a larger model.

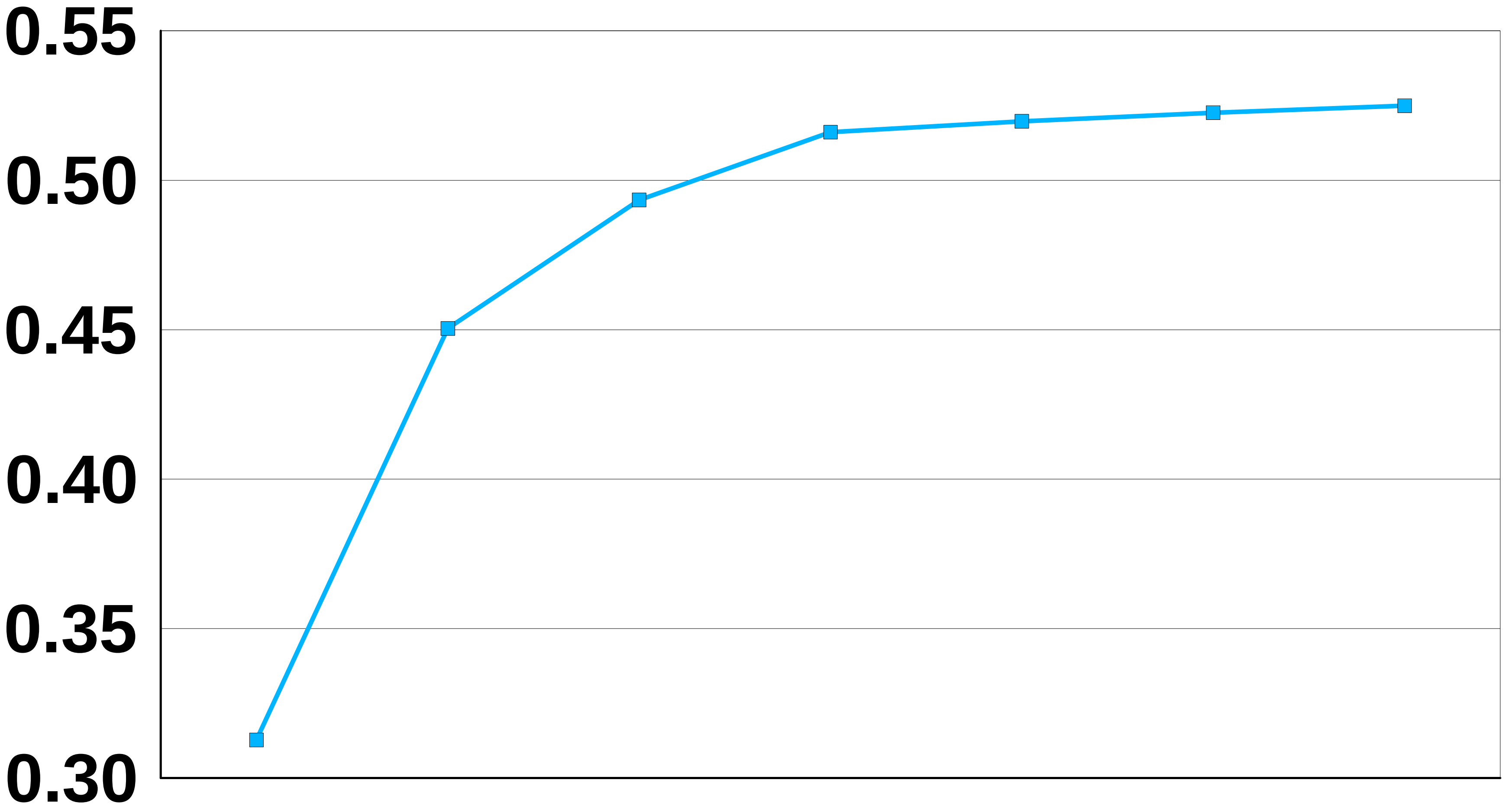
```
proc reg data=SENIC;
model InfRisk = LtofStay Age CulRatio XRay NoBeds
                Census Nurses Services
          / selection=stepwise sle=.5 sls=0.5;
```

- and got these results

```
        Variable  Variable  Number    Partial     Model
Step    Entered   Removed   Vars In   R-Square   R-Square   C(p)    F Value    Pr>F
  1     CulRatio                1      0.3127     0.3127   41.5161    50.49  <.0001
  2     LtofStay                2      0.1377     0.4504   13.3525    27.57  <.0001
  3     Services                3      0.0430     0.4934    5.9368     9.25  0.0029
  4     XRay                    4      0.0227     0.5161    2.9592     5.07  0.0263
  5     Nurses                  5      0.0036     0.5197    4.1729     0.80  0.3731
  6     NoBeds                  6      0.0029     0.5226    5.5431     0.64  0.4260
  7     Age                     7      0.0023     0.5249    7.0440     0.50  0.4795
```

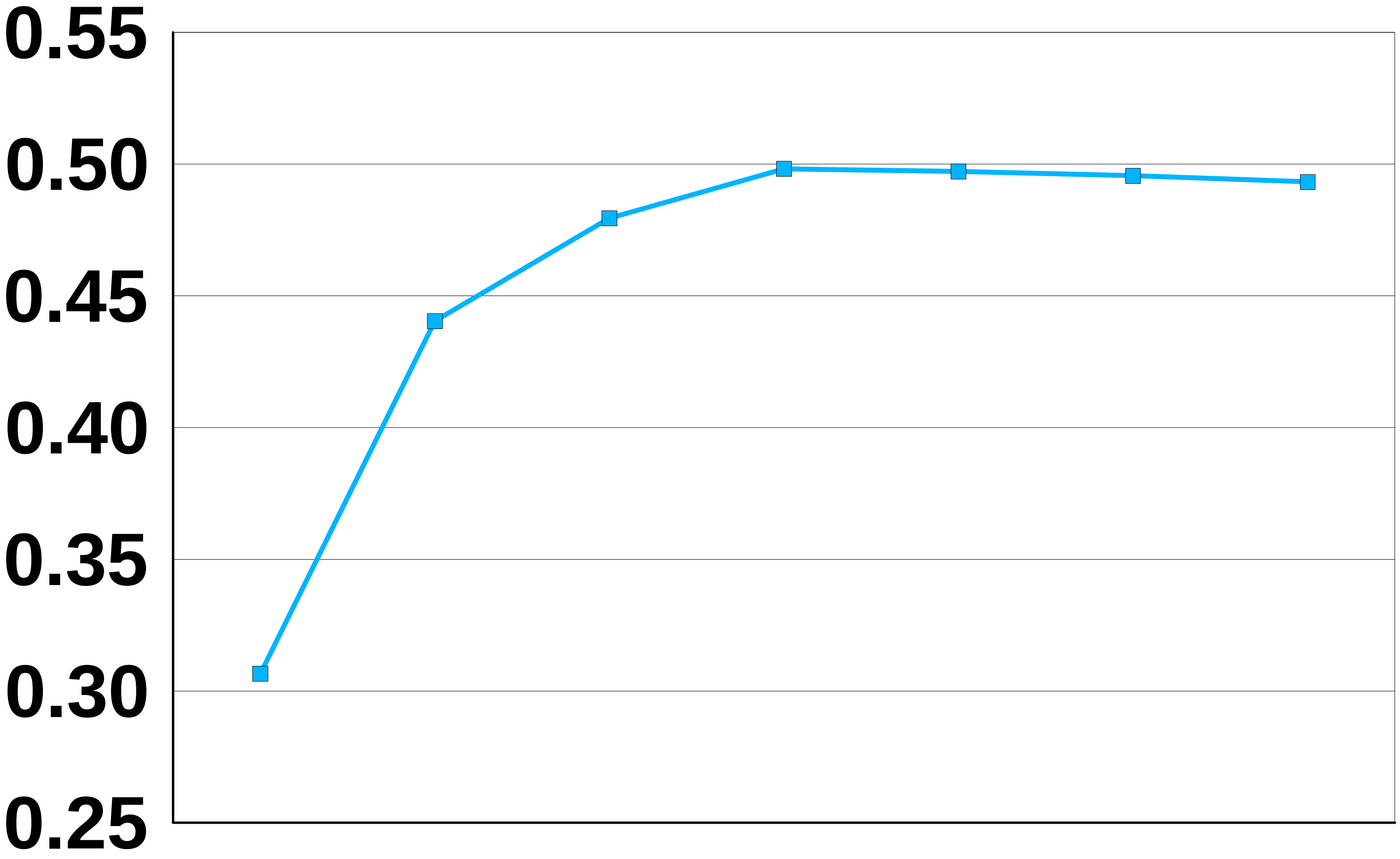
Stepwise Selection (*continued*)

- Plotting the ordinary R^2 value gives the following graph (always larger).



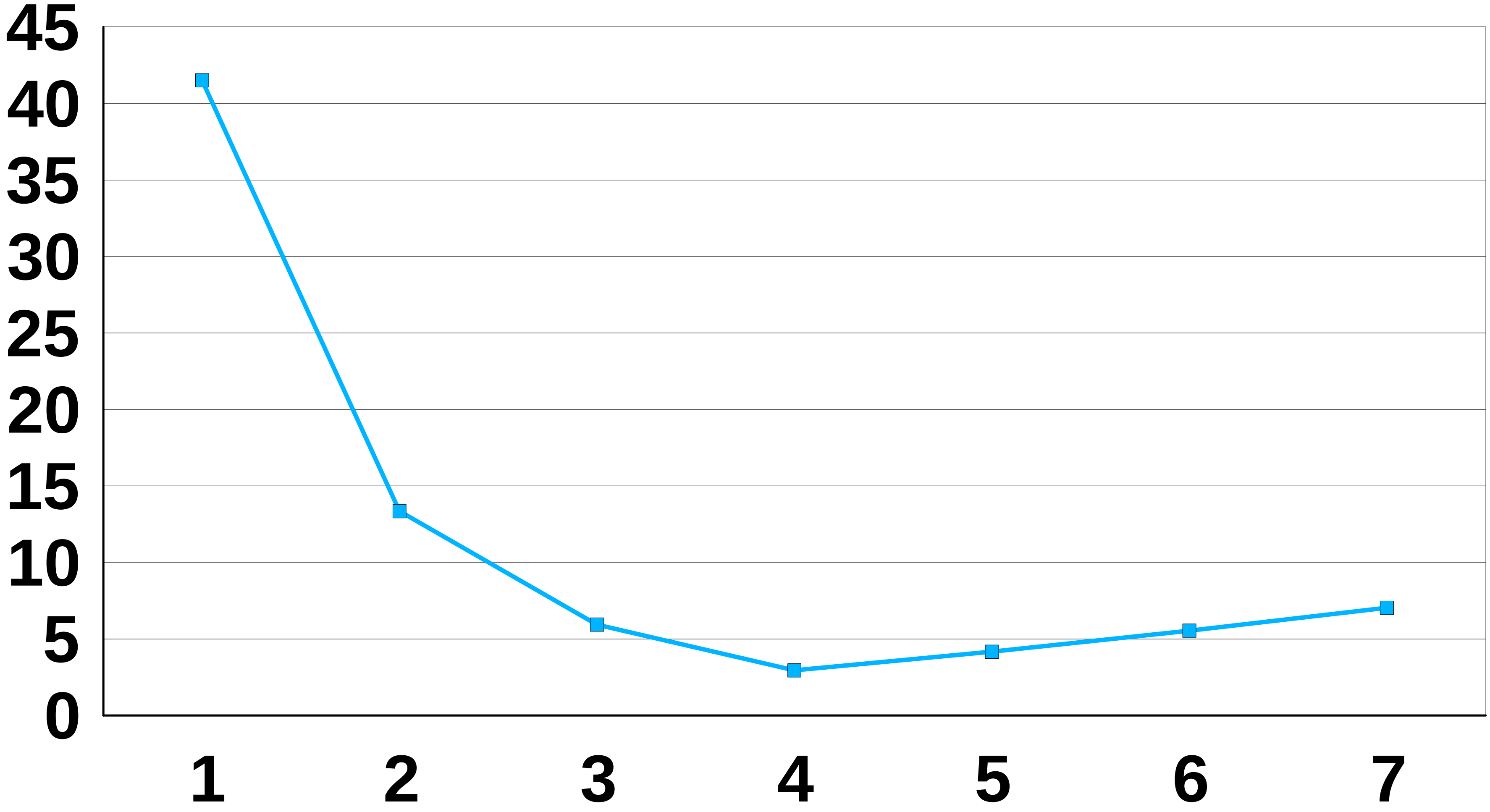
Stepwise Selection *(continued)*

- The adjusted R^2 value goes down a little after the 4th variable. I calculated these in EXCEL.



Stepwise Selection *(continued)*

- Mallow's $C_{(p)}$ statistic is supposed to indicate the "best" model when $C_{(p)}$=p.



Stepwise Selection *(continued)*

- Unfortunately, Mallow's $C_{(p)}$ statistic depends on the Full model being a pretty good model with no multicollinearity.
- This is not always true, and as we know it is probably not true for this example.
-

Multicollinearity of the Reduced Model

- Question! Did the stepwise selection and the resulting reduced model cure our "multicollinearity" problems?
- I reran the reduced model with options to get the collinearity diagnostics.

```
Variable      Variance Inflation
Intercept                      0
LtofStay                 1.35777
CulRatio                 1.28045
XRay                     1.33266
Services                 1.15566
```

Multicollinearity of the Reduced Model *(continued)*

- The mean is less than two, and no values even reach the mean, much less the criteria of 10.
- The highest condition number was only 16, well below the criteria of 30.
- The model selected with stepwise regression clearly has no multicollinearity problems.

Multicollinearity of the Reduced Model *(continued)*

- **The reduced model has 4 significant variables, more than the 3 significant variables in the full model, and not the same variables that were significant.**
- **It also has an R^2 = 51.61% while the full model had R^2 = 52.51.**
- **The simpler model with nearly the same R^2 value is most likely a very superior model.**

The R^2 selection option

- **There is one other "variable selection" option that is very interesting. It is quite different from the stepwise selection model.**
- **Suppose that you are going to fit a model with a number of variables, lets call them a, b, c, d, e, and f.**
- **What happens if stepwise selection chooses one set of variables, but for some reason you prefer a different set?**

The R^2 selection option *(continued)*

- **For example, if you feel that variables a, b & d should be the best variables, and stepwise selects a, b and e. How much better is this model than the one that you feel is best. Or suppose that variable d is inexpensive and easy to measure while c is expensive and difficult. If you use d instead of c, how much do you loose?**
- **We will examine the R^2 selection option.**

The R^2 selection option *(continued)*

- **This procedure will show you the best models, not just one, but several.**
- **It will also show you how good larger (more variables) and smaller (fewer variables) models might be.**
- **The major criteria here is the value of R^2, which is something of a limitation.**

The R^2 selection option *(continued)*

- **To request the procedure ask for model options "selection=rsquare". I also included the options "start=3 stop=6 best=8";**
- **This instructs SAS to start with 3 variable models, go up to 6 variables and show me the best 8 models for each number of variables.**
-

The R^2 selection option *(continued)*

- **As requested, the RSQUARE selection option first produces the best 8 3-factor models (plus intercept).**

```
Number    R-square  Variables in Model
in Model
     3  0.49340010  LTOFSTAY CULRATIO SERVICES
     3  0.48523075  LTOFSTAY CULRATIO NURSES
     3  0.47356336  LTOFSTAY CULRATIO NOBEDS
     3  0.47347050  LTOFSTAY CULRATIO CENSUS
     3  0.46955655  LTOFSTAY CULRATIO XRAY
     3  0.46300398  CULRATIO XRAY SERVICES
     3  0.46191250  CULRATIO XRAY CENSUS
     3  0.45384242  CULRATIO XRAY NURSES
```

The R^2 selection option *(continued)*

- **And then the best 4-factor, 5-factor, etc.**

```
Number in   R-square   Variables in Model
  Model
      4   0.51613081   LTOFSTAY CULRATIO XRAY SERVICES
      4   0.51023237   LTOFSTAY CULRATIO XRAY NURSES
      4   0.50002851   LTOFSTAY CULRATIO XRAY CENSUS
      4   0.49971593   LTOFSTAY CULRATIO XRAY NOBEDS
      4   0.49556642   LTOFSTAY CULRATIO NURSES SERVICES
      4   0.49556459   LTOFSTAY AGE CULRATIO SERVICES
      4   0.49348607   LTOFSTAY CULRATIO NOBEDS SERVICES
      4   0.49341314   LTOFSTAY CULRATIO CENSUS SERVICES
```

- **The best model we found was a 4-factor model. Here we can check for alternative 4-factor models.**

The R^2 selection option *(continued)*

- **Note that frequently very little is lost by replacing one or two variables with different variables, often less than a few percentage points on the R^2 value.**
- **The other variables may be more interpretable, more reliably measured, cheaper and easier to measure, or have some other advantage.**

Other Regression Topics

- As mentioned earlier, the intercept for our last problem was not very meaningful (when all X_i equal zero we have no beds, no nurses, a length of stay of zero days, etc.)
- This is not an uncommon problem. In studying marine organism, for example, a salinity of zero is not a marine environment, a temperature of zero is not liquid and a depth of zero is not wet.

Other Regression Topics *(continued)*

- So, if you want to plot your data on one of the X_i values, what do you do. If you just extract the intercept and slope of interest, you are essentially setting all other X_i equal to zero. This can lead to unreasonable values of Yhat even if you do not show the intercept.
- Yhat = $b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + b_4x_{4i}$
- Yhat = $b_0 + b_1x_{1i} + b_2(0) + b_3(0) + b_4(0)$
- Yhat = $b_0 + b_1x_{1i}$

Other Regression Topics *(continued)*

- In order to do a plot of Y_j and $Yhat_j$ on a single X_{ij} value, it is best to set the other X_{ij} values to their mean value.
- $Yhat = b_0 + b_1X_{1j} + b_2X_{2j} + b_3X_{3j} + b_4X_{4j}$
- $Yhat=b_0 +b_1X_{1j} +b_2(\bar{X}_2) +b_3(\bar{X}_3) +b_4(\bar{X}_4)$
- $Yhat=[b_0+b_2(\bar{X}_2)+b_3(\bar{X}_3)+b_4(\bar{X}_4)]+ b_1X_{1j}$
- Since all $b_i\ \bar{X}_i$ are "constant", the part in brackets is a new "intercept", b'_0, then $Yhat = b'_0 + b_1x_{1i}$

Other Regression Topics *(continued)*

- For the final 4-factor model, If I wanted to plot our observed and predicted SENIC values on Length of Stay (with a meaningful range of values) I would get the following results.
-

Other Regression Topics *(continued)*

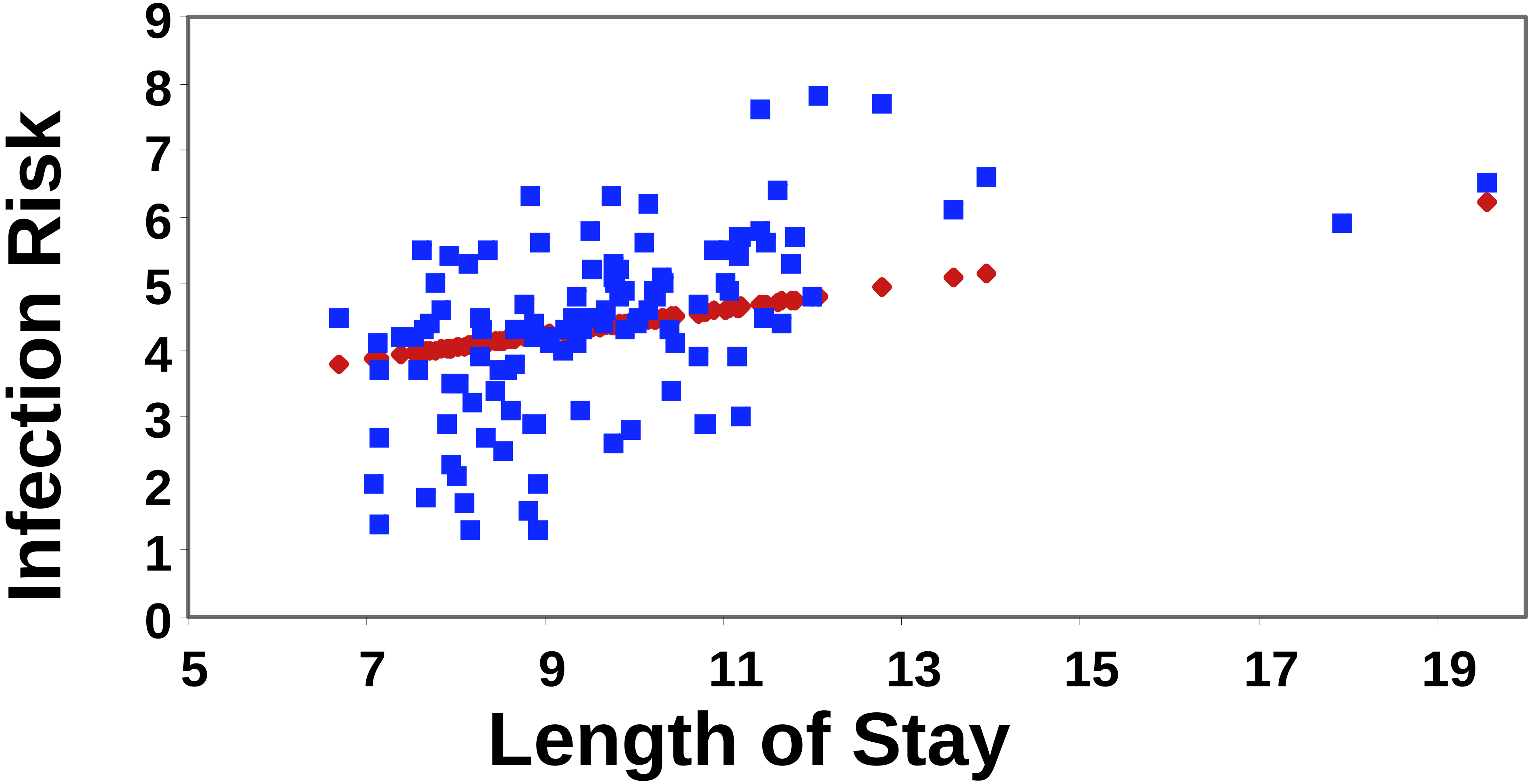
Variable	Parameter Estimate	Means	Constants	SLR
INTERCEP	-0.06358059		-0.06358059	2.53702
LTOFSTAY	0.18841053			0.18841
CULRATIO	0.04644573	15.79	0.733513715	
XRAY	0.01205242	81.63	0.983818779	
SERVICES	0.02046537	43.16	0.88327088	
		Sum	2.537022785	

Other Regression Topics *(continued)*

- Notice the change in intercept, it is no longer negative, suggesting that even for a very short stay in the hospital (near zero time) there is still a positive risk of infection. This seem more reasonable.
- Now lets look at the plot of the adjusted model.

Other Regression Topics *(continued)*

Observed and Predicted infection risk

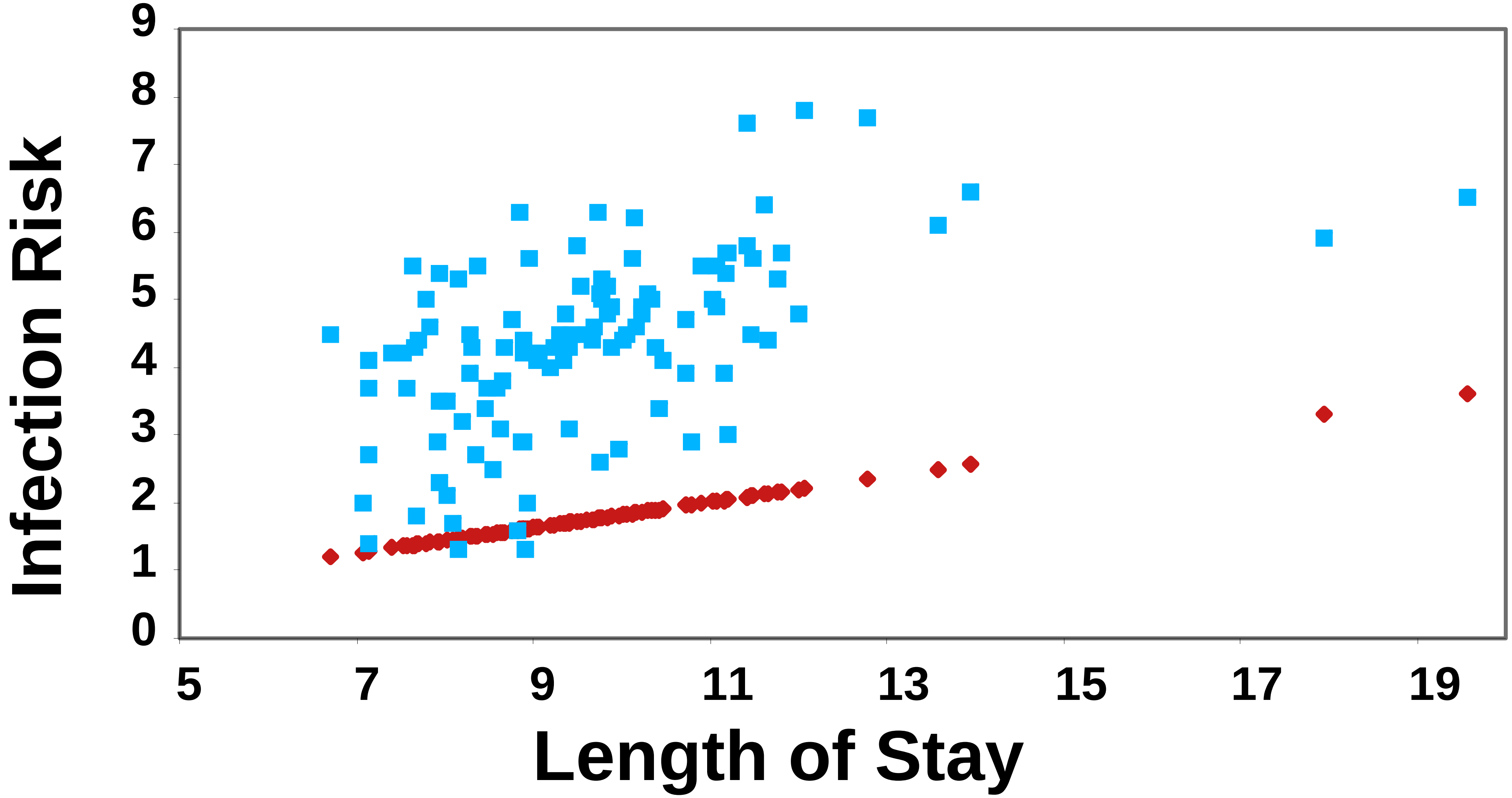


Other Regression Topics *(continued)*

- The predicted values line up nicely, as we would expect for a simple linear regression, and could be connected to fit a line.
- Also, though the origin is not shown, the intercept of 2.5 looks reasonable.
- The next graph used the full model slope and intercept to get the predicted values. All other X_i values are essentially set to zero.

Other Regression Topics *(continued)*

Observed and Predicted Infection Risk



Other Regression Topics *(continued)*

- **Obviously this line does not fit well, and it's negative intercept is too low.**
- **There appears to be a great deal of scatter, but remember we are looking at the Y variable on only one X variable. There are 3 other significant independent variables doing their share to explain the variation.**

Cause & Effect

- I reiterate, you cannot prove cause and effect with correlation or regression. Cause and effect are "proved" with a controlled experiment.
- However, once proved relationships can be quantified with regression,
- and a good correlation may prove to be a useful predictive tool even where there is no cause and effect.

Linear combinations

- Regression is a linear combination. It is linear because the terms are additive.
- There are some properties of linear combinations that are useful not only for regression, but for other applications as well.
-

Linear combinations *(continued)*

- **Take the linear combination**
  - **$A_i = aX_i + bY_i + cZ_i$**
  - **$Var(A_i)=a^2*Var(X_i) + b^2*Var(Y_i) + c^2*Var(Z_i)$ +2*Covariances**
  - **$Var(A_i)=a^2\sigma^2_{Xi} + b^2\sigma^2_{Yi} + c^2\sigma^2_{Zi} +2(ab\sigma_{Xi,Yi} + ac\sigma_{Xi,Zi} + bc\sigma_{Yi,Zi})$**
  - **Unless the variables are independent, in which case the covariances may be assumed to be zero.**

Linear combinations *(continued)*

- **For our variance calculation purposes in Multiple regression**
  - **we need not consider the covariance among observations because they are independent,**
  - **We need not consider the covariance among Yhat$_i$ and e$_i$ because they are independent**
  - **We DO NOT consider the parameter estimates of a multiple regression independent, and we use the covariance estimates from the analysis.**

Linear combinations *(continued)*

- **Other applications,**
- **In a two-sample t-test, and later on in Analysis of variance, if you want to test an hypothesis between two or more independent estimates like,**
  - ► **H_0: $\mu_1 = 0.5\mu_2$ or $\mu_1 - 0.5\mu_2 = 0$**
- **We note that since these are independent, the variance for this t-test will be**
  - ► **$Var(\mu_1) + 0.5^2 Var(\mu_2)$**

Linear combinations *(continued)*

- **Linear combinations also are used in sampling.**
- **If random sampling is done on a heterogeneous population, the heterogeneity will cause a large variance. If the population is broken into smaller, more homogeneous, units the variance of each of the units will be smaller.**

Linear combinations *(continued)*

- **The overall variance is then calculated by summing the individual variances (multiplied by the square of the coefficients). Since the units are sampled independently no covariance is needed.**
- **For an example, with calculations, see "Linear combinations" under the EXST7005 notes.**

General Linear Hypothesis Test

- This test is relatively easy in view of what we know about extra SS.
- In this test we can examine the addition of any variable or group of variables to a model. The model without the variables is called the Reduced model. The model with the additional variables we want to test is called the Full model.

General Linear Hypothesis Test *(continued)*

- **For example, suppose we had previously seen a study such as the SENIC hospital study that had two variables only, Length of stay and average age of the patient (X_1 and X_2). We want to jointly test the other 6 variables to see if they add anything JOINTLY to the model.**
- **To do this we would calculate the extra SS= $SSX_3, X_4, X_5, X_6, X_7, X_8 \mid X_1, X_2$**

General Linear Hypothesis Test *(continued)*

- **We fit the Reduced model and the Full model.**
- **Full model results: dfError = 104, SSE=95.63982**
- **Reduced model results: dfError=110, SSE=141.99965**

General Linear Hypothesis Test *(continued)*

- Then we set up the table below to test the difference.

Source	d.f.	SSE	MSE	F	P>F
Reduced model	110	141.9997			
Full model	104	95.6398			
Difference	6	46.3598	7.7266	8.4020	0.00000020
Full model	104	95.6398	0.9196		

General Linear Hypothesis Test *(continued)*

- **In this case we see that the difference is highly significant, indicating that the amount of variation described by the omitted variables is significantly different from zero. At least one of these variables would be a useful addition to the model.**
-

Multiple Regression Summary

- Although the observation diagnostics are similar between SLR and MLR, there a number of new diagnostics for variables.
- There is also a new problem (multicollinearity) that needs to be addressed. Don't forget, or underestimate, this problem.

Multiple Regression Summary *(continued)*

- The assumptions for MLR are basically the same as for SLR.
- Most diagnostics on assumptions and model adequacy are similar (normality, curvature, etc.).
- We have partial residual plots (which could have been done for SLR) as a new diagnostic tool.

Multiple Regression Summary *(continued)*

- **Extra SS are important to understanding the various types of SS, and the General Linear Test.**

Multiple Regression Summary *(continued)*

- **You should now be able to interpret the parameter estimates provided by SLR or MLR, and use most of the diagnostics produced by SAS to determine variable "importance", evaluate observations and determine of the model is adequate and if the assumptions are met!**
- **Congratulations.**