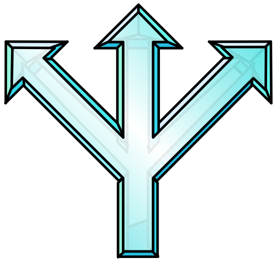
Statistical Techniques II

EXST7015

Multiple Regression Summary & Example 2



Objectives

- Example 2 is a larger multiple regression. Before looking at output, lets consider our objectives.
  - Objectives can vary.
  - You are probably interested in testing for relationships (actually "partial" correlations) between the various independent variables and the dependent variable.
- Why?

Objectives *(continued)*

- On the one hand, you may be interested in the effect of each and every variable included in the analysis.
  - If this is the case, you are probably particularly interested in the parameter estimates.
  - And, you will probably be interested in confidence intervals on these parameter estimates, or on testing them against certain hypothesized values.

Objectives *(continued)*

- Since you are interested in all of the variables, you are probably not interested in removing any variables from the model.
  - Our first example may have been this type of analysis. We wanted to examine the effect of 3 soil phosphorus components, and would present the results for all 3, even though only one was significant, because all 3 are of interest.

Objectives *(continued)*

- On the other hand, you may not be interested in all of the variables. You may not know which variables are important, and your objective may be to determine which ones are important.
  - In this case you may want to keep and discuss all of the variables.
  - Or you may want to select the important variables and present them in a smaller (reduced) model with only significant variables.

Objectives *(continued)*

- **Our second model is more likely to be of this type. There are 8 variables, and most likely not all are important. Our objective here is probably to determine which ones are *correlated* to the dependent variable. We may start with a full model, but will then probably reduce the model to some subset of significant variables.**

Example 2

- This example is from Neter, Kutner, Nachtsheim and Wasserman. 1996. Applied Linear Statistical Models. Irwin Publishing Co.
- It is a sample various hospitals. Data was taken for the "Study of the Efficacy of Nosocomial Infection Control" (SENIC Project).
- Each observation is for one hospital. There were 113 observations.

Example 2 *(continued)*

- The variables are as follows;
  - Identification number
  - Average length of stay of all patients (in days)
  - Average age of all patients (years)
  - Average Infection risk for the hospital (in percent)
    - This was taken as dependent variable
  - Routine culture ratio to patients without symptoms to patients with symptoms

Example 2 *(continued)*

- **The variables (continued);**
  - **Routine chest x-ray ratio for patients with and without symptoms of pneumonia**
  - **Average number of beds (during study)**
  - **Med School Affiliation (1=Yes, 2=No)**
  - **Region NE=1, NC=2, S=3, W=4**
  - **Average no. of patients during study**
  - **Average no. of nurses**
  - **Percent of 35 potential service facilities that were actually provided by the hosp.**

Example 2 *(continued)*

- Infection risk was used as the dependent variable.
- Eight other variables were considered independent variables.
- The two class (categorical, group) variables were omitted from the analysis (Med school affiliation and Region).
- Identification number was not used.

Example 2 *(continued)*

- **See the computer output. Selected output was omitted as trivial or not of interest.**
- **Data list -**
  - **A partial listing of the data is provided on the handout. A complete listing is available in the SAS program and output.**
  -

Interpretation and evaluation

- The PROC REG statement is given at the bottom of the first page.
  - Only one new option is included here, the COLLIN option on the model statement.
  - We have seen all of the other output down through the PROC UNIVARIATE,
  - and the PROC GLM on the last page.
  - But not the Stepwise regressions.

Interpretation and evaluation *(continued)*

- **So, which variables are important, if any?**
- **Are there problems with Multicollinearity, and what do you do about it anyway?**
- **Are there any problems with the observations?**
  - **Y_i variable outliers?**
  - **X_i variable "outliers"?**
  - **Influential observations?**

Interpretation and evaluation *(continued)*

- **Are the assumptions for linear regression met? Normality? Independence? Homogeneity of variance? Are the X_i variables measured without error?**
- **And what are our objectives? If we want find the best model for predicting infection risk, do we need all 8 variables? How do we go about removing the ones we don't want?**

Example 2

- Turn to the computer output.
- In the Program,
  - Note that a label statement is present, but has been turned into a comment to prevent labels in the program. This was done only to create a smaller output.

Example 2 *(continued)*

- The proc print on the handout has had observations removed to conserve space. All observations are available in the SAS output listing and in the program.
- Two variables are "class" or "indicator" variables. These were not included here, but use of this type of variable will be discussed later under "Analysis of Covariance".

Example 2 *(continued)*

- **Some output has been deleted to save space. This is information that I considered very simple, or information that we will not cover this semester,**
  - **Descriptive Statistics - simple, you should already know these**
  - **Uncorrected Sums of squares and Crossproducts - redundant, repeats information in the X'X section.**

Ex 2 - Correlation

- Correlation section - here we can get out first hints at possible multicollinearity. Look for the larger correlations (>0.9, maybe even >0.8).
  - There are a number of correlations at this level, especially among the variables NOBEDS, CENSUS, NURSES, and to a lesser extent SERVICES.
  -

Ex 2 - Correlation *(continued)*

- **The highest correlation is 0.98 between NOBEDS and CENSUS.**
- **There are several other correlations above 0.90. Correlation with SERVICES are not as high, but there are several of them (remember the MULTI in MULTIcollinearity)**
- **The available evidence suggests quite strongly that there could be multicollinearity problems.**

Ex 2 - X'X, X'Y and Y'Y

- The next section is the "Model Crossproducts X'X X'Y Y'Y" section. It contains all SS and CP for all X_{ij} and Y_i variables.
  - Since there are 8 independent variables, the first 8 rows and 8 columns are the X'X matrix. The last column on the right is the X'Y vector, except for the value in the last row and column (lower, right corner) which is the Y'Y value.

Ex 2 - X'X, X'Y and Y'Y *(continued)*

- **Test yourself, could you find the following values?**
  - ►**The SS for $X_{LTOFSTAY}$, or X_{NURSES}. The values are 10928.3862 and 5563895 respectively.**
  - ►**Find the crossproducts of AGE with the dependent variable, INFRISK. This value is 26196.13,**
  - ►**Find the value for Y'Y, the SS of the dependent variable. It is 2344.41.**

Ex 2 - $(X'X)^{-1}$, B vector and SSE

- **The next section is titled "X'X Inverse, Parameter Estimates, and SSE". It contains the X'X inverse matrix needed to calculate the Variance-Covariance matrix. It also has the B vector (regression coefficients) and the SSE. However, we can find these values elsewhere.**
- **For this section, know how the $(X'X)^{-1}$ relates to the Variance-Covariance matrix.**

Ex 2 - Matrix solution

- **We did not cover matrix algebra, but I expect you to know something about these.**
  - **As with the SLR, we need all SS and CP from the variables to do a multiple regression. I expect you to know where these are.**
  - **The Variance-Covariance matrix is key to most variance calculations. Know where this is and where it comes from.**

Ex 2 - ANOVA table

- The next section is the "Analysis of Variance" table.
  - It is of somewhat less interest in multiple regression, because it has all variables tested jointly. We probably want to test the variables individually.
  - However, you should be thoroughly familiar with this section and its contents. It has changed little since we discussed SLR.

Ex 2 - ANOVA table *(continued)*

- Analysis of Variance

```
                 Sum of      Mean
Source    DF    Squares    Square  F Value  Prob>F
Model      8  105.74000  13.21750   14.373  0.0001
Error    104   95.63982   0.91961
C Total  112  201.37982

    Root MSE   0.95896     R-square   0.5251
    Dep Mean   4.35487     Adj R-sq   0.4885
    C.V.      22.02053
```

- The only value we have not discussed on this table is the Adjusted R^2 value. Discussion of that value comes soon.

Ex 2 - Parameter Estimates

- This is a key section since it deals with the evaluation of variables.
  - First we look at the variables and their tests against zero. Recall that any variable that does not differ from zero contributes little to the model.
  - Also recall that the regression coefficients are interpretable parameter estimates. How much does infection risk increase for each additional day in the hospital?

Ex 2 - Parm. Est. *(continued)*

- Parameter Estimates

```
              Parameter    Standard   T for H0:
Variable  DF   Estimate       Error Parameter=0  Prob>|T|
INTERCEP   1  -0.747255  1.20762993      -0.619    0.5374
LTOFSTAY   1   0.176931  0.06905830       2.562    0.0118
AGE        1   0.016214  0.02225515       0.729    0.4679
CULRATIO   1   0.046993  0.01074904       4.372    0.0001
XRAY       1   0.012037  0.00548557       2.194    0.0304
NOBEDS     1  -0.001447  0.00271027      -0.534    0.5945
CENSUS     1   0.000728  0.00347158       0.210    0.8343
NURSES     1   0.001906  0.00175316       1.087    0.2794
SERVICES   1   0.016280  0.01018895       1.598    0.1131
```

Ex 2 - Parm. Est. *(continued)*

- **Things to note.**
  - **The intercept is -0.74. What would this mean? Well, first of all it is not different from zero, so it does not probably mean much.**
  - **But then, the intercept is for a length of stay of zero days, for an average age of zero, in a hospital that, on the average, does zero cultures and zero X-rays, has no beds, no patients, no nurses and no services. This is not a very interesting parameter, rather theoretical.**

Ex 2 - Parm. Est. *(continued)*

- Things to note.
  - The increase risk of infection is 0.177 for each additional day of stay in the hospital. Since the dependent variable is in percent, this means that each day has about a 0.177% increase risk of infection. This variable is "significant" (H_o: $\beta_{ltofstay} = 0$ would be rejected).
  -

Ex 2 - Parm. Est. *(continued)*

- **Things to note (continued).**
  - **Note that CULRATIO is even more significant (P>|t| , 0.001 compared to 0.0118 for Ltofstay). However the value is much smaller ($b_{culratio}$=0.047 compared to $b_{ltofstay}$=0.177).**
  - **How can this be? The culture ratio variable has a much smaller standard error. It is estimated with greater precision.**
  - **So the actual size of the value of the b_i is not what is important.**

Ex 2 - Parm. Est. *(continued)*

- Interestingly, in addition to length of stay the two other significant variables are culture ratio (Routine culture ratio to patients without symptoms to patients with symptoms) and XRay (Routine chest x-ray ratio for patients with and without symptoms of pneumonia).
- Maybe this increase in infection risk is because more infections are found on a routine basis.
- No other significant variables are indicated at this time.

Ex 2 - Parm. Est. *(continued)*

- **Additional statistics of interest in the Parameter Estimates section**

```
                                                Standardized
Variable  DF    Type I SS  Type II SS       Estimate
INTERCEP   1  2143.030177    0.352108     0.00000000
LTOFSTAY   1    57.305110    6.036475     0.25221461
AGE        1     2.075060    0.488094     0.05394755
CULRATIO   1    31.457347   17.576777     0.35868492
XRAY       1     3.847649    4.427818     0.17382256
NOBEDS     1     6.516419    0.262196    -0.20812617
CENSUS     1     0.174357    0.040435     0.08347342
NURSES     1     2.016416    1.087193     0.19797721
SERVICES   1     2.347646    2.347646     0.18454925
```

Ex 2 - Type I and Type II SS

- **For this analysis we have no interest in the TYPE I SS.**
- **We also have little interest in the TYPE II SS because we will use the t-tests of the slopes for statistical evaluation.**
  - **It is interesting to note, however, that all variables had TYPE II SS smaller than the TYPE I SS,**
  - **and that the pattern is different from the regressions coefficients (it will match the t-value pattern of course).**

Ex 2 - Standardized Regression Coefficients

- We might be interested in the standardized regression coefficients. We saw earlier that the Ltofstay variable had the largest regression coefficient, but not the largest t value. Note that the Std. Reg. Coeff. match the t-value pattern for the 3 most significant variables, not the raw reg. coeff.

Ex 2 - Std. Reg. Coeff. *(continued)*

- However, the standardized regression coefficients do not match the t value pattern for all variables!
- XRay was the third significant variable, with the third largest t value, but three variables that were not significant (smaller t-values) had larger absolute values for the std. reg. coeff. (NOBEDS, NURSES and SERVICES). What's up? More later.

Ex 2 - Partial R^2 values

- Another set if statistics for evaluation is the various R^2 values.

```
                          Squared        Squared        Squared        Squared
                     Semi-partial        Partial   Semi-partial        Partial
Variable  DF        Corr Type I    Corr Type I   Corr Type II   Corr Type II
INTERCEP   1           .              .              .              .
LTOFSTAY   1     0.28456232     0.28456232     0.02997557     0.05936955
AGE        1     0.01030421     0.01440266     0.00242375     0.00507754
CULRATIO   1     0.15620903     0.22153115     0.08728172     0.15524912
XRAY       1     0.01910643     0.03480702     0.02198740     0.04424825
NOBEDS     1     0.03235885     0.06107540     0.00130200     0.00273400
CENSUS     1     0.00086581     0.00174046     0.00020079     0.00042261
NURSES     1     0.01001300     0.02016338     0.00539872     0.01123980
SERVICES   1     0.01165780     0.02395864     0.01165780     0.02395864
```

Ex 2 - Partial R^2 *(continued)*

- **Since we use the Type II SS for this problem we would probably be interested in the Squared partial correlation Type II.**
- **Note that these values follow pattern more similar to the t values than the standardized regression coefficients. The three largest match the 3 significant variables, and in the same order. However, the match is not perfect across all variables.**

Ex 2 - Parameter Estimate Section

- **We have 3 statistics we can use to interpret the contribution or importance of the variables. I consider the t-test to be the key test, and the std. reg. coeff and partial R^2 type values are ancillary statistics.**

Ex 2 - Variance Inflation Factor (VIF)

- Now lets consider multicollinearity. Recall variance inflation. Is it possible that some other variable is important, but does not show up because it is in competition with another variable for the SS? Or that two multicollinear variables have inflated variance and do not have significant t-tests because of this?
- Lets see.

Ex 2 - VIF *(continued)*

- VIF values

```
                                         Variance
Variable  DF      Tolerance             Inflation
INTERCEP   1      .                    0.00000000
LTOFSTAY   1    0.47122355             2.12213501
AGE        1    0.83280592             1.20075995
CULRATIO   1    0.67841753             1.47401851
XRAY       1    0.72771534             1.37416369
NOBEDS     1    0.03005773            33.26931229
CENSUS     1    0.02881692            34.70183289
NURSES     1    0.13774001             7.26005447
SERVICES   1    0.34228836             2.92151333
```

Ex 2 - VIF *(continued)*

- Two VIF values (NOBEDS and CENSUS) are greater than 10, a clear indication of problems.
- A second variable (NURSES) is greater than 7. I would consider this a possible problem as well.
- We have talked about this problem, but have not discussed what to do about it.
- So, what do we do?

Ex 2 - Multicollinearity

- **Solving Multicollinearity problems.**
  - **First we know that with multicollinearity problems we have potentially wide fluctuations in the regression coefficients.**
  - **And that we have inflated variances.**
  - **Since the problems are caused by two or more correlated variables, one obvious solution is to leave off some of the variables that are correlated.**

Ex 2 - Multicollinearity *(continued)*

- **This is often the easiest and best solution. Fit a reduced model that has some subset of the original variables and where correlated variables are reduced.**
- **A second solution is to get more data. Sometimes additional data will bring out the differences in the variables and reduce the correlation.**

Ex 2 - Multicollinearity *(continued)*

- **A third solution is "Ridge Regression". This is an interesting regression, an example of where a statistician may seek biased estimates in order to reduce variance.**
- **However, it can be hard to interpret the results or to decide exactly how much bias needed. I consider it a last resort, but one I have used.**
- **We will not discuss this option in detail.**

Ex 2 - Multicollinearity *(continued)*

- **We will apply the first and best option to this example (reducing the number of variables).**
- **When we reduce a model like this, we do it one variable at a time. We will use "Stepwise regression" to do this.**
- **This section will come when we finish with the rest of Example 2 in PROC REG.**

Ex 2 - Variance-Covariance Matrix

- Covariance of Estimates
  - This section was deleted, but it is simply the $(X'X)^{-1}$ matrix multiplied by the MSE. It is available in the SAS program output. You are responsible for knowing what is in this matrix and how it relates to the parameter estimate variances.
- Correlation of Estimates (deleted)
  - You are not responsible for these

Ex 2 - Sequential Parameter Estimates

- **Recall that when multicollinearity exists, reg. coeff can fluctuate greatly. In this section we look to see if there are large fluctuations, or if the parameter estimates are stable.**

Ex 2 - Sequential Parameter Estimates *(continued)*

- **Examining each variable,**
  - **LTOFSTAY is quite stable. The biggest change is from 0.239 to 0.170 when NOBEDS enters. Maybe hospitals with fewer beds move patients out faster, but I don't think this is a very serious change.**
  - **AGE changes a lot when CULRATIO enters, even changing sign. I think there might be is a problem here.**

Ex 2 - Sequential Parameter Estimates *(continued)*

- Examining each variable,
  - CULRATIO and XRAY are very stable.
  - NOBEDS changes a lot with CENSUS and again (including a sign change) with NURSES. There is probably a problem here.
  - Census changes with NURSES.
  - The rest are pretty stable.

Ex 2 - Sequential Parameter Estimates *(continued)*

- **So the Sequential Parameter Estimates provide an additional indication of problems, and some indication of which variables are affected by which others.**

Ex 2 - Collinearity Diagnostics

- I have included one additional section that you have not seen before. It is the "Collinearity Diagnostics" section.
  - This is produced by the "collin" option on the model statement.
  - We will be concerned only with the first 3 columns, and particularly the last few numbers in the third column.

Ex 2 - Collinearity Diagnostics *(continued)*

- **Multicollinearity diagnostics is a statistic based on eigenvalues and eigenvectors. This technique extracts abstract axes that describe variation among the variables. If all variables are perfectly uncorrelated, each would have an eigenvalue of 1. If perfectly correlated the first value would be equal to p, and the rest would be zero.**

Ex 2 - Collinearity *(continued)*

```
Collinearity Diagnostics
                          Condition
Number   Eigenvalue           Index
     1      7.92221         1.00000
     2      0.70667         3.34824
     3      0.23449         5.81248
     4      0.04727        12.94588
     5      0.03554        14.93049
     6      0.02878        16.59008
     7      0.01657        21.86586
     8      0.00546        38.09407
     9      0.00301        51.29072
```

```
Collinearity Diagnostics
                             Condition    ------------------------Proportion of Variation-----------------------
  Number      Eigenvalue         Index      Intercept        LtofStay        CulRatio            XRay        Services
       1         4.66066       1.00000        0.00125         0.00124         0.00872         0.00176         0.00405
       2         0.21630       4.64186        0.00937         0.00468         0.85226         0.00149         0.03500
       3         0.07914       7.67398        0.02688         0.00983         0.02687         0.10305         0.81528
       4         0.02612      13.35827        0.11951         0.30253         0.05132         0.84526         0.14359
       5         0.01778      16.19088        0.84299         0.68173         0.06084         0.04843         0.00208
```

Ex 2 - Collinearity Diagnostics *(continued)*

- **The relative variation accounted for can be evaluated by looking at the ratio of the first eigenvalue (√) to the others (√), particularly the last one(s). If the ratio is greater than about 30, multicollinearity is indicated.**
  -

Ex 2 - Collinearity Diagnostics *(continued)*

- **In our example the highest values are 51 and 38, so there is considerable multicollinearity. There is a further suggestion that at least two variables must be removed to remove the multicollinearity.**
-

Ex 2 - Multicollinearity Summary

- **Collinearity Diagnostics are a very good tool to evaluate multicollinearity, perhaps the best.**
- **VIF is also a very good tool, and is the most popular.**
- **Simple correlations are good, but may miss higher order correlations ("multi").**
- **Sequential b_i values are useful, but somewhat subjective.**

Ex 2 - More diagnostics

- You are not responsible for the sections titled,
  - Consistent Covariance of Estimates (deleted)
  - Test of First and Second Moment Specification (deleted)
- These have been deleted from the handout.

Ex 2 - Observation Diagnostics

- Not all of the 113 observations are present in the handout, but they are available in the SAS output list.
- I have tried to include some of the more interesting observations, and have placed underscores to indicate where sections have been deleted.

Ex 2 - Observation Diagnostics *(continued)*

- The first columns are the value of Y_i and the predicted value of Y_i. You are responsible for understanding these, along with the residual (the difference between these two values). These have not changed from SLR.
- You are not responsible for the Std Err Predict or the Std Err Residual. These are estimates of standard deviations and have been adjusted by h_{ii} values.

Ex 2 - Observation Diagnostics *(continued)*

- You are responsible for the confidence intervals, Upper and Lower 95% MEAN and Upper and Lower 95% PREDICT. These are confidence intervals for the regression line and for individual points respectively.

$$Y_i = b_0 + b_1 X_i + e_i$$

$$Y_i = \hat{Y} + e_i$$

Ex 2 - Observation Diagnostics *(continued)*

- **You are responsible for the Studentized residual, and perhaps more important the Deleted studentized residual (RSTUDENT).**
  - **In this example there are 113 obs. and 104 d.f. in the model. The critical t value for one observation is 1.983, and Bonferroni adjusted for 113 (-1) obs. is 3.627.   The largest value is obs. # 53 at 3.0267. This does not exceed the Bonferroni adjusted value.**

Ex 2 - Observation Diagnostics *(continued)*

- The mean h_{ii} value should be 9/113 = 0.0796, so no hat value should exceed 2*0.08 = 0.16. Hospital # 8 has a value of 0.3516, and # 112 has a value of 0.4289.
- These hospitals should be examined for conformity to the study objectives, but high hat values are not an indication of problems, only of unusual values.

Ex 2 - Influence Diagnostics

- **You are not responsible for the column titled Cov Ratio.**
- **The remaining 3 diagnostics of interest are influence diagnostics. The criteria for evaluation are,**
  - **DFFITS : 2*sqrt(p/n) = 0.564**
  - **DFBetas : 2/sqrt(n) = 0.188**
  - **Cook's D : F dist. (1-α, 9, 104 d.f.)**
    - **F(30th percentile) = 0.708**
    - **F(50th percentile) = 0.933**

Ex 2 - Influence Diagnostics *(continued)*

- **Evaluation of influence in Ex 2.**
  - **Using Cook's D - No values exceed our criteria here. There are no influential statistics as evaluated with Cook's D. Remember that this is across all reg. coeff. and there are a lot of reg. coeff.**
  - **Using DFBetas - Our criteria here is about 0.2, and several observations have changes above this value. The highest are**
    - **Obs # 53 : Intercept = -0.6207**

Ex 2 - Influence Diagnostics *(continued)*

- Evaluation of influence in Ex 2.
  - Using DFBetas - Our criteria here is about 0.2, and several observations have changes above this value. The highest are
    - # 53: Int = -0.62, Age = 0.47, NoBeds=1.0981, Census=-0.8620
    - #8: Int=-1.43, CulRatio=-0.94, XRay=0.32, NoBeds=-0.87, Census=0.65
    - And other obs. have values exceeding the criteria

Ex 2 - Influence Diagnostics *(continued)*

- **Evaluation of influence in Ex 2.**
  - **Using DFFits - Our criteria here is about 0.6, and several observations have changes above this value. The highest are**
    - **#8 DFFits = -1.4251**
    - **#11 = -0.9745,**
    - **#53 = 1.3383**
    - **#112 = -1.1413**

Ex 2 - Influence Diagnostics *(continued)*

- **So there are numerous influential observations in this analysis, particularly #8. #53. #112 and #11.**
- **But remember that "influential" is not the same as "outlier". There may be no problem with these points.**
- **However, #8 and #112 had unusual sets of X_i values, so they should be examined carefully.**

Ex 2 - Influence Diagnostics *(continued)*

- **But there were no outliers as judged with RStudent, so we may not have any real serious problems.**
- **The largest RStudent values was 3.0267 for observation #53, so this record should also be examined.**

Ex 2 - Partial Residual Plots

- **These are "scatter plots" of the Y variable adjusted for all X_i except one plotted on that X_i adjusted for all other X_i.**
- **I used these to get across the concept that not only are the Y_i adjusted for each X_i, but the X_i are also adjusted for each other.**

Ex 2 - Partial Residual Plots *(continued)*

- **Beyond this these are used more like "scatter plots" than "residual plots".**
  - **We can look for curvature, nonhomogeneous variance, etc.**
  - **If they appear to represent random scatter about zero it is because the variable does not contribute anything to the model, not because it is a "residual plot".**

Ex 2 - Ordinary Residual Plots

- Included in the SAS output.

```
Full Model with diagnostics                           Plot of e*YHat.  Legend: A = 1 obs, B = 2 obs, etc.
        ,                                                                                 A
        ,
        ,
      2 ˆ                                        A                     A
        ,                          A
        ,                                          A   A    A                 A
        ,                                                                          A                                       A
        ,                                                           A
        ,                                     A A                   A                           A
      1 ˆ                                                          A           A
        ,                                                           A     A          A A
   R    ,                            A            A A                                    A  A
   e    ,                                             C      AA                   A       A
   s    ,                                                 A     AA           A  A A             A                       A                            A
   i    ,                                             A           AA AA  A     AA     A              A            A
   d  0 ˆffffffffffffffffffffffffffffffffffffffffffffffffffffffffAAfffffAfffffAfffffffffffffAffffffffffffAfffffffffffffffffffffffffffffffffffffffffffff
   u    ,                                  A   A     AA      A              A A  A     B     A
   a    ,                                                 A       A      A                                                             A
   l    ,        A          A               AA                      A   A    A      A      A         A
        ,                                           A                A                   AA
        ,                               A          A   A AA                       A    A                   A
     -1 ˆ             A                            A                                                                                     A
        ,                              A                       A        A
        ,                    A                A                      A                         A                A
        ,                        A                                                                                                        A
        ,                                  A    A
        ,                                                                                                                              A
     -2 ˆ
        ,                                                                         A
        ,

        Šffˆffffffffffffˆffffffffffffˆffffffffffffˆffffffffffffˆffffffffffffˆffffffffffffˆffffffffffffˆffffffffffffˆffffffffffffˆffffffffffffˆfffffffff
          2.0          2.5          3.0          3.5          4.0          4.5          5.0          5.5          6.0          6.5          7.0
                                                                      Predicted Value of InfRisk
```

Ex 2 - Ordinary Residual Plots

```
 SENIC database from NKNW 1996 (Appendix C)
Full Model with diagnostics
                                                      Plot of E*YHAT.  Legend: A = 1 obs, B = 2 obs, etc.

        |                                                                                 A
        |
        |
      2 +                                        A                     A
        |                          A
        |                                           A   A    A                 A
        |                                                                           A                                       A
        |                                                           A
        |                                     A A                   A                           A
      1 +                                                          A           A
        |                                                           A     A          A A
   R    |                            A            A A                                    A  A
   e    |                                             C      AA                   A       A
   s    |                                                 A     AA           A  A A             A                       A                            A
   i    |                                             A           AA AA  A     AA     A              A            A
   d  0 +--------------------------------------------------------AA-----A-----A-------------A------------A--------------------------------------------
   u    |                                  A   A     AA      A              A A  A     B     A
   a    |                                                 A       A      A                                                             A
   l    |        A          A               AA                      A   A    A      A      A         A
        |                                           A                A                   AA
        |                               A          A   A AA                        A    A                   A
     -1 +             A                            A                                                                                     A
        |                              A                       A        A
        |                    A                A                      A                         A                A
        |                        A                                                                                                        A
        |                                  A    A
        |                                                                                                                              A
     -2 +
        |                                                                        A
        |
        ---+------------+------------+------------+------------+------------+------------+------------+------------+------------+------------+--------
          2.0          2.5          3.0          3.5          4.0          4.5          5.0          5.5          6.0          6.5          7.0
                                                                      Predicted Value of INFRISK
```

Ex 2 - SAS Version 8

- A few notes on SAS versions 6 and 8.
- The "config.sas" (V6) or "Sasv8.cfg" (V8) file in the SAS directory has an option to change between character types.
- Some of the styles are fine if you have the SAS monospace font installed, but lines and hash marks show up as odd letters if you do not.
-

Ex 2 - SAS Version 8 *(continued)*

- I usually use the last of the SAS options for character sets. It does not require the SAS fonts.
- /* This is the OEM character set   */
- /* -FORMCHAR ³ÄÚÂ¿ÃÅ´ÀÁÙ+=|-/\<>*   */
- /* This is the ANSI character set (for SAS Monospace font and ANSI Sasfont) */
- /* -FORMCHAR ‚ƒ„…†‡ˆ‰Š‹Œ+=|-/\<>*   */
- /* This is the ANSI character set   */
- -FORMCHAR |----|+|---+=|-/\<>*

Ex 2 - SAS Version 8 *(continued)*

- The new SAS version 8 defaults to a high resolution graphic instead of the older character graphics.
- I often use BATCH SUBMIT and do not want high resolution graphics.
- Include the new "LINEPRINTER" option to get character graphics in SAS version 8.
-

Ex 2 - Residual Analysis with PROC UNIVARIATE

- This is an important procedure for evaluating residuals, especially for the assumption of normality.
- The Shapiro-Wilk test values are
  - W:Normal   0.985827   Pr   0.8069
- These results would lead you to FAIL to reject the hypothesis of normality. The results are consistent with a normal distribution.

Ex 2 - Residual Analysis with PROC UNIVARIATE *(continued)*

- The plots lead to the same conclusion.

```
Stem Leaf                     #  Boxplot                        Normal Probability Plot
 24 6                        1     0         2.5+                                                  *
 22                                             |                                                 ++
 20 2                        1     |            |                                               *+
 18 7                        1     |            |                                             *+
 16 836                      3     |            |                                         **+*
 14 6989                     4     |            |                                       ***+
 12 2                        1     |            |                                      *++
 10 32458                    5     |            |                                    ***
  8 797                      3     |            |                                  +**
  6 0899279                  7     |            |                                +**
  4 5623557                  7  +-----+         |                              +***
  2 024614567899            12  |     |         |                            ***
  0 238803455679            12  *--+--*      0.1+                         ***
 -0 5541943                  7  |     |         |                       ***
 -2 88322110                 8  |     |         |                     ***
 -4 9864310762210           13  +-----+         |                  ****
 -6 99327                    5     |            |                 **
 -8 55864421                 8     |            |              ****
-10 291                      3     |            |             **
-12 7424                     4     |            |          ***
-14 3720                     4     |            |        **
-16 23                       2     |            |     **+
-18 5                        1     |            |   *++
-20                                |            |  ++
-22 2                        1     |        -2.3+*+
    ----+----+----+----+                         +----+----+----+----+----+----+----+----+----+----+
Multiply Stem.Leaf by 10**-1                         -2        -1         0        +1        +2
```

Ex 2 - Residual Analysis with PROC UNIVARIATE *(continued)*

- We would also check for outliers, and again see no great problems. Obs # 53 is too large, but only one out of 113, so not entirely unexpected.
- This is consistent with our observations from the RStudent values.
-

Ex 2 - GLM output

- There is not much that is new here. Just note that,
  - PROC GLM does the same analysis as PROC REG, either can be used.
  - PROC GLM provides Type I SS and TYPE III SS by default, and provides tests of these sums of squares.
  - The tests of the Type III SS (or TYPE II SS) are identical to the t-tests of the regression coefficients.

Conclusion

- One additional section of Multiple Regression will discuss variable selection and a few other topics to complete the Multiple Regression section.
- This will be followed by sections on curvilinear regression and logistic regression to complete the Regression part of the course.