DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

MORE ON MULTIPLE REGRESSION

I. AGENDA:
   A. Multiple regression
      1. Categorical variables with more than two categories
      2. Interaction and specification
   B. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition, Chapter 12, pages 449 to 462.

II. SUBSTANTIVE PROBLEM:
   A. What explains voting for H. Ross Perot, an independent candidate for the United States presidency in 1992 and 1996? Some hypotheses:
      1. "Disguised conservatism": It's possible that he was an attractive alternative to conservatives or that self-identified independents really are more conservative than neutral. If so, then places that supported Reagan most solidly in 1984 tended to give Perot his largest totals
      2. "Decline in party strength": One also wonders if support for Perot is not partly a (negative) function of attachment to political parties. The period after 1970, say, has been called an age of "party decline" and voter demobilization. So in areas where party loyalty to either party is weak, Perot, the independent, may have greatest appeal. More precisely, where split ticket voting--a majority for, say a Republican candidate for president and for a Democratic for Senate--Perot will draw his greatest support.
      3. "Regional-cultural factors": Perot seemed more popular in the West than the East. Furthermore, some western states have a history of political "independence."
   B. Using the data set "Perot," which is available on the web site, we can (very crudely) investigate some of these ideas.
      1. The units are counties in four states.
   C. Operational definitions:
      1. Dependent variable (**Y**): percent of total votes cast that were for Perot in 1992.
      2. Independent variable (**Z**): Percent of votes for Reagan in 1984.
         i. This is a crude indicator of political conservatism.
         ii. It's not totally unreasonable because political scientists feel that many political attitudes are relatively enduring. So an area that was conservative in 1984 have roughly the same characteristic in 1992.
      3. Independent variable (**X**): State. We can use "state" as surrogate or

indicator for cultural and political climate.

i. In the data they are coded as:

1 New Jersey
2 Colorado
3 Maryland
4 Oregon

III. CATEGORICAL INDEPENDENT VARIABLES:

A. We have already seen how to translate a two-category variable into a dummy variable.

1. In particular, suppose a variable has two categories or levels, A and B. Then we create two "dummy" variables, X_1 and X_2, as follows:

> X_1 = 1 if unit has characteristic A (is in category A); 0 otherwise.
> X_2 = 1 if unit has characteristic B (is in category B); 0 otherwise

2. But, we only use <u>one</u> of these variables in the analysis because the second one is redundant.

i. That is, once we know the value of X_1, we can predict exactly the corresponding value of X_2. Hence, the variables are perfectly related and keeping both in a model would convey no extra information.

B. Generalization: suppose a categorical variable has **K** categories or levels. Then we create **K - 1** dummy variables. (The Kth would be redundant or unnecessary as in the case above.)

1. As an example let's use "state" as a "predictor" of vote for Perot. Since "state" has four "classes," we have to form K - 1 = 4 - 1 = 3 dummy variables as follows:

> X_1 = 1 if New Jersey, 0 otherwise
> X_2 = 1 if Colorado, 0 otherwise
> X_3 = 1 if Maryland, 0 otherwise

2. The state of Oregon becomes the reference category: counties in Oregon are coded 0 on all three dummy variables.

3. An aside: there is nothing wrong or misleading with treating "state" as a variable. In fact, doing so allows us to quantify (to an extent) the notion of state culture or political climate.

i. Later we will create a region variable by combining various states.

C. Iincidentally, the newer versions of MINITAB commands create coded variable automatically. But you can also use the “code” procedure as well. Suppose state (X) has been stored in column 3 with the following codes:

1 New Jersey
2 Colorado
3 Maryland
4 Oregon

i. We need to change the numbers (1, 2, 3, 4) into dummy variables because these integers are really only convenient labels, not actual measures of anything. Here's how we might proceed:
   1) Open the file and then go to the worksheet.
   2) You can type in these commands at the MINITAB prompt (assume the state variable is stored in c3):

```
MTB > code (2,3,4) to 0 in c3 put in c4
<Note: c4 now contains only 0's and 1's>
MTB > code (1,3,4) to 0 in c3 put in c5
MTB > code (2) to 1 in c5 put back in c5
<Note: c5 now contains only 0's and 1's>
MTB > code (1,2,4) to 0 in c3 put in c6
MTB > code (3) to 0 in c6 put in c6
<Note: c6 now contains only 0's and 1's>
# Pay careful attention to the coding.
```

1. You can look at the results in the data window to make sure they are what you want.
   i. You should see 1's, 2's, 3's, and 4's in c3. In c4 0's will correspond to the 2's, 3's and 4's in c3. The other columns will be coded similarly. And use **Tally** to make sure that your variables have only 0 and 1.
   ii. Once again, pay attention to the coding shown above. If it doesn't make sense, look in the data window to see what is happening at each step.

D. As we saw in Class 14, dummy variables can be entered into a regression model just as any others can.
   1. Moreover, interpret the numerical values of the coefficients in the usual way:
      i. A one unit change in X_1 (with the other variables held constant) leads to $\hat{\beta}_1$ units change in Y
   2. Alternatively and more meaningfully, use the strategy described in the last class:
   3. Write the estimated general model:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

i. Now consider the situation when $X_1 = X_2 = X_3 = 0$. (Example: counties within Oregon.) The model reduces to:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \\ &= \hat{\beta}_0 + \hat{\beta}_1(0) + \hat{\beta}_2(0) + \hat{\beta}_3(0) \\ &= \hat{\beta}_0 \end{aligned}$$

ii. For the reference category (the units having 0's on the X's) the model reduces to $\hat{\beta}_0$, which is the average value of the response variable for these units. That is,

$$\hat{\beta}_0 = \bar{Y}_{(ref\ group)}$$

4. The other $\hat{\beta}'s$ are the "effects" of being in a particular category (e.g., being in a particular state.

5. Once again, substitute the values of the X's to see how to interpret the equations. Example consider those units coded 1 on X_1 (e.g., counties in New Jersey). The estimated model reduces to:

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2(0) + \hat{\beta}_3(0) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \\ &= \hat{\beta}^* \end{aligned}$$

i. The $\hat{\beta}^*$ is the mean of the units in the first category of the categorical variable (e.g., counties in New Jersey.) $\hat{\beta}_1$ represents the "change" or effect of "moving" from the reference category (e.g., being in Oregon) to the first category (e.g., think of a county's moving to New Jersey from Oregon).

E. It is perhaps easier to see this interpretation by looking at the actual data. Here is the estimated model for the Perot data:

$$\hat{Y} = 26.9 - 9.36X_{NJ} - 1.10X_{CO} - 9.74X_{MD}$$

1. If we let $X_1 = X_2 = X_3 = 0$, which is the case for counties in Oregon, the predicted value of Y is 26.92. This is the mean percent for Perot among Oregon's 36 counties.
2. The effect of being in New Jersey is to reduce this mean by 9.36 (the value of $\hat{\beta}_1$) to 17.56 (i.e., 26.92 - 9.36 = 17.56). Hence, the mean vote for Perot in New Jersey is 17.56.
3. If this is not clear do as suggested above: place the values for New Jersey counties in the estimated equation:

$$\begin{aligned}\hat{Y}_{NJ} &= 26.92 - 9.36(1) - 1.10(0) - 9.74(0)\\ &= 26.92 - 9.36\\ &= 17.56\end{aligned}$$

4. One can interpret the other values similarly.

F. The means for Perot for the four states are:

State	Mean
NJ	**17.56 = 26.92 - 9.36**
CO	**25.82 = 26.92 - 1.10**
MD	**17.17 = 26.92 - 9.94**
OR	**26.92 = 26.92**

Table 1: Mean Vote for Perot

1. The mean for each state is equal to the average value of the reference category minus the effect (i.e., the $\hat{\beta}_k$) of being in a particular state. (There is no coefficient for Oregon so its value is just its mean.)
2. The data seem to fit this model reasonably well as can be determined by the regression results.
   i. See the table on the next page.

Predictor	Coef	Stdev	t-ratio	p
Constant	26.9194	0.7409	36.34	0.000
C4	-9.362	1.221	-7.67	0.000
C5	-1.1020	0.9287	-1.19	0.237
C6	-9.740	1.171	-8.32	0.000

s = 4.445 R^2 = .471

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	3	2459.44	819.81	41.49	0.000
Error	140	2766.26	19.76		
Total	143	5225.70			

Table 2: Regression Results for Perot Data

ii. Note that both the F for the overall model and the individual coefficients are significant; the multiple R is reasonably large.

iii. Still, one wonders if the model can't be improved--that is, if we can't achieve greater understanding of Perot's support--by adding an additional variable and interaction.

IV. ANALYSIS OF VARIANCE:

A. First, it's worth pausing to see the relationship between regression analysis with dummy variables and a procedure called analysis of variance (ANOVA, for short).

1. See Agresti and Finlay, ***Statistical Methods for the Social Sciences***, 3r edition, pages 439 to 449.

B. Imagine two or more populations that can meaningfully be measured on some variable, Y. Each of these populations will have a mean value of Y, μ_j for J = 1, 2, 3,...to however many populations there are.

1. A natural question is: are these μ's all equal or do they differ.
2. Suppose we were comparing two separate populations of counties (say, those in the East versus Western ones) in terms of support for Perot. Would the means be the same or different.

C. The ANOVA null hypothesis:

$$H_0: \quad \mu_1 = \mu_2 = \ldots \mu_K = \mu$$

where μ *is the "grand" or overall mean*

1. The research or alternative hypothesis is:

H_A: $\mu_j \neq \mu_k$

for at least some j and k.

2. The null hypothesis is tested with an F test: one computes an observed F by obtaining the ratio of the mean square for groups to the mean square for errors. This F has K - 1 and N - K - 1 degrees of freedom, where K is the number of populations. It turns out that this is exactly the same as the F obtained by getting the ratio of mean square for regression to mean square for residual.
   i. Since we are not dwelling on the point, I won't prove it. But it is easily seen by comparing the ANOVA and regression tables and results.

D. MINITAB
1. If you want to use ANOVA directly, just use MINITAB or SPSS's analysis of variance procedures.
   i. Here's an example.
   ii. See Agresti and Finlay, Statistical Methods for other examples of computer printout.

```
                ANALYSIS OF VARIANCE ON PEROT
SOURCE      DF         SS         MS         F         p
PEROT        3     2459.4      819.8     41.49     0.000
ERROR      140     2766.3       19.8
TOTAL      143     5225.7

                                      INDIVIDUAL 95% CI'S FOR MEAN
                                      BASED ON POOLED STDEV
 LEVEL      N      MEAN   STDEV --+---------+---------+---------+----
     1     21    17.557  4.817       (----*----)
     2     63    25.817  5.277                              (--*-)
     3     24    17.179  3.598     (----*---)
     4     36    26.919  2.817                               (--*---)
                                 --+---------+---------+---------+----
POOLED STDEV =    4.445           16.0      20.0      24.0      28.0
```

Table 3: Analysis of Variance Results for Perot Data

2. The means equal the values obtained above when we substituted into the

regression equation.

3. The observed F, which is compared to an critical F, $F_{(\ ,K-1,N-K-1)}$, and is used to test the null hypothesis, gives the same value as when we regress K dummy variables on Y. Hence, the $F_{obs} = 41.49$ in this example is the same as the value obtained from the regression analysis.
    i. Keep in mind when doing regression with a categorical variable having K categories, the degrees of freedom associated with the regression sum of squares is K - 1; similarly, ANOVA on a set of K populations has K - 1 degrees of freedom.
    ii. Of course, <u>if</u> additional variables are added to the regression equation, as done next, the degrees of freedom for regression will change. (We lose a degree of freedom for every variable entered into the equation.)

V. INTERACTION REVISITED:
  A. Suppose we are interested in the first hypothesis mentioned on page 1.
    1. It states, in essence, that the vote for Perot will be related to vote for Reagan.
  B. Look carefully at the following figure:
    1. It shows vote for Perot by vote for Reagan for each of the states.
      i. You can create such a “multiple plot” by going to the graph menu, then to the regions menu, and indicating that all the plots are to be drawn or place on one page.

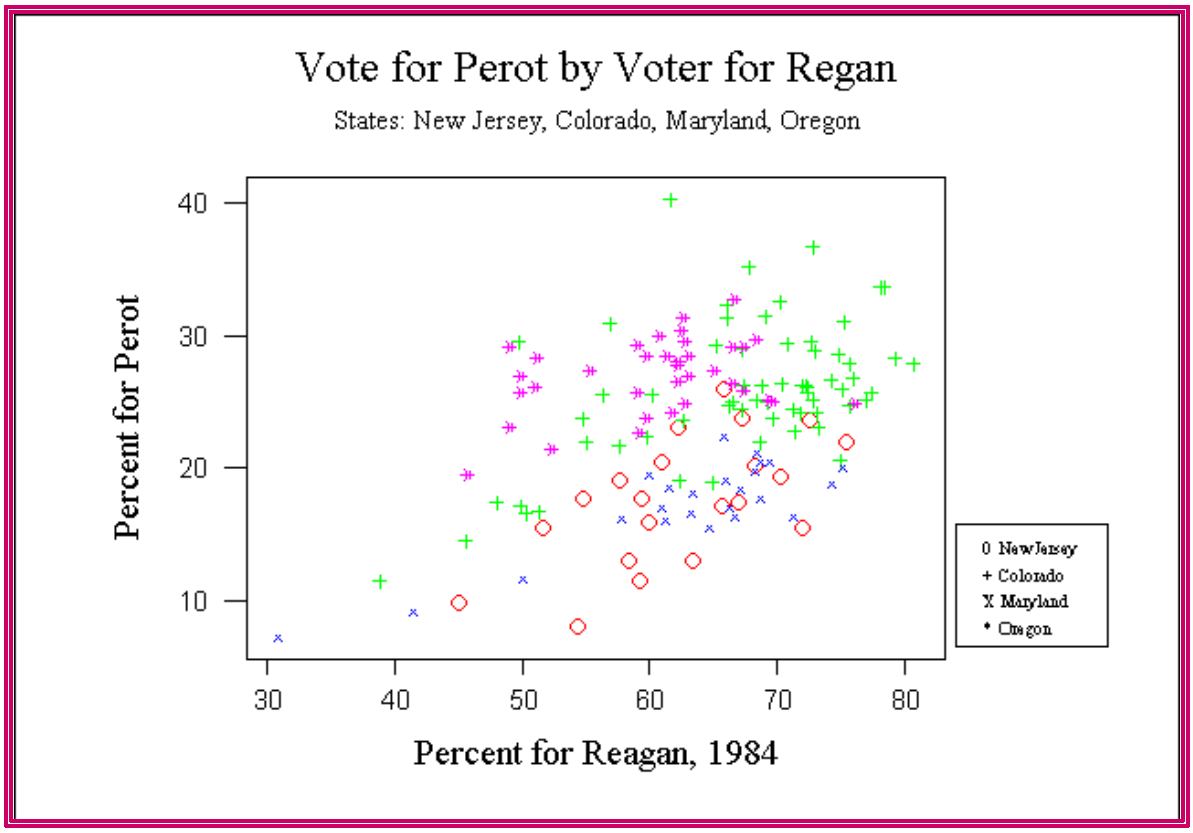

Figure 1: Vote for Perot By Reagan Vote and State

C. It may not be apparent at first but if you look carefully, you will notice that the symbols representing Oregon counties seem to follow a much more horizontal pattern than do the rest. This suggest that the relationship between Perot and Reagan vote is weaker there than in the other states.
   1. But this is simply specifying a relationship: stating under what conditions it holds.
   2. Here my claim is that the relationship between the Y and Z differs in Oregon.
   3. This is just another way of saying that there is interaction between Perot voting and Reagan support, the nature of the relationship being affected by X, a variable representing perhaps "state political culture."

D. To test this idea, let's create an interaction variable:
   1. That is, we need an indicator or dummy variable that distinguishes counties in Oregon from those in the other states.
   2. Actually, we can just multiply the Oregon dummy variable by the quantitative independent variable of interest.
   3. In MINITAB we can create the variable by multiplying Z (the quantitative independent variable) by the dummy variable, X_4, for Oregon.

4. In the calculator or mathematical expressions dialogue box indicate that the new variable is to be store in, say, column 10, and the it is the result of multiplying the column for Oregon (c4) by the column forZ, say c7.
   i. In the work session this could be expressed as: **let c10 = c2*c7**.
   ii. Imagine what happens when we do this: column c7 contains 0's for all the state counties except those in Oregon. Hence, column 10, which contains the interaction variable, will have mostly 0's and those values of vote for Reagan for each of Oregon's counties.

E. Now we can regress Perot vote on 1) vote for Reagan, 2) the dummy variable for Oregon, and 3) the interaction term. The model is thus:

$$E(Y) = \hat{\beta}_0 + \hat{\beta}_{Reagan}Z + \hat{\beta}_{Oregon}X + \hat{\beta}_{interaction}W$$

1. Note again that Z is the percent for Reagan, X is the dummy variable for Oregon (1 if county is in Oregon, 0 otherwise) and W is the interaction (Z*X), which will equal 0 when X = 0 and Z when X = 1.
2. The OLS estimates are:

$$E(Y) = -4.157 + .4Z + 23.388X - .278W$$

F. To interpret the numbers and understand their political or theoretical significance, make the usual substitutions. First, look at the counties not in Oregon, the ones coded 0. The equation reduces to:

$$\begin{aligned} EY) &= -4.157 + .4Z + 23.388(0) - .278(0) \\ &= -4.157 + .4Z \end{aligned}$$

1. The equation for Oregon counties, on the other hand, differs because now X = 1 and hence W = Z(1) = Z. Thus, for Oregon the equation is:

$$\begin{aligned} E(Y) &= -4.157 + .4Z + 23.388(1) - .278(1)Z \\ &= (-4.157 + 23.88) + (.4 - .278)Z \\ &= 19.23 + .127Z \end{aligned}$$

2. Look at the β's--remember they are partial or controlled coefficients--for the two categories of states:
   i. In Oregon the relationship between Perot voting and Reagan support is only .127, whereas in the other states it is about three times larger.
   ii. Perhaps the next figure will help clarify the situation.

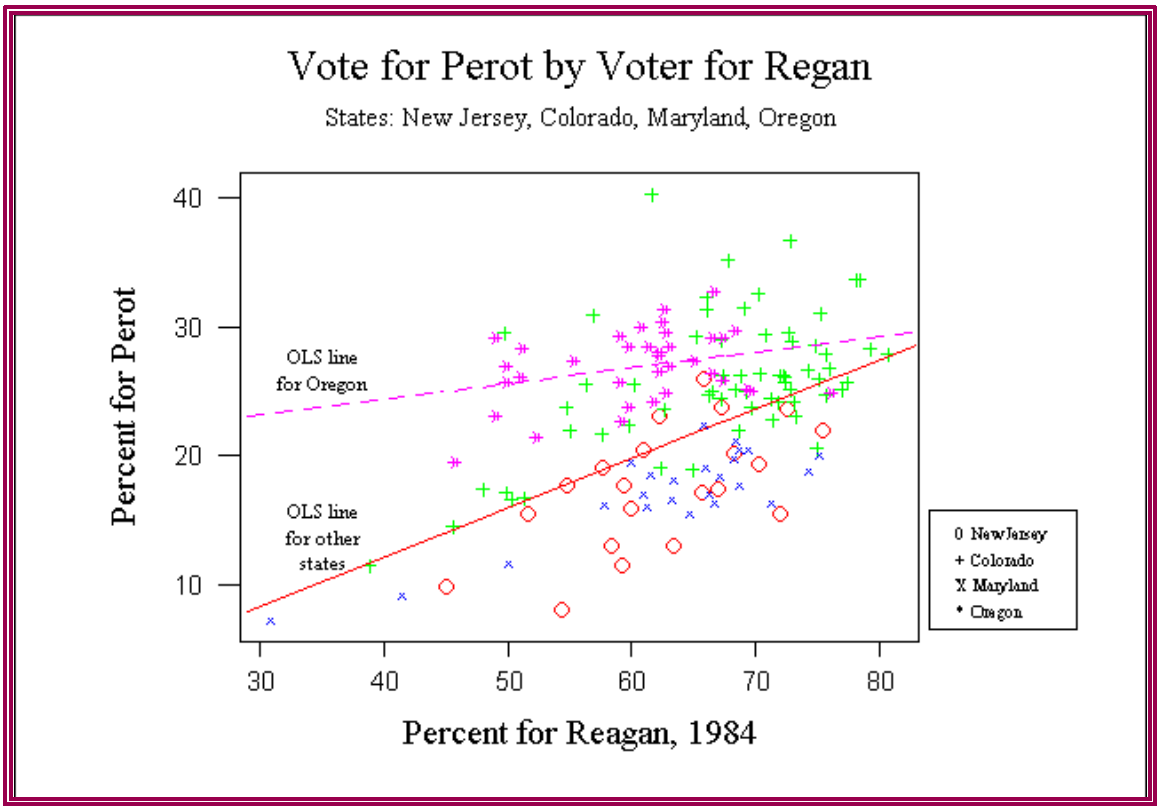

Figure 2: Perot By Reagan With OLS Lines

3. Note that the least squares lines differ, which is what we would expect if interaction is present.

G. Is the interaction term statistically significant? Here are the tests of the overall models with and without the interaction term.
   1. First, the reduced model, the one without the interaction variable.

Predictor	Coef	Stdev	t-ratio	p
Constant	-1.295	3.006	-0.43	0.667
C2	0.36140	0.04552	7.94	0.000
C7	6.3043	0.9424	6.69	0.000

s = 4.773 R^2 = .385

Table 4: Regression With No Interaction Term

2. The R^2 is about .38, a value that can be compared with the one for the complete model, the model containing interaction.
3. Here it (the complete model) is:

Predictor	Coef	Stdev	t-ratio	p
Constant	-4.157	3.222	-1.29	0.199
C2	0.40525	0.04888	8.29	0.000
C7	23.388	7.614	3.07	0.003
C10	-0.2784	0.1232	-2.26	0.025

s = 4.705 R^2 = .407

Table 5: Regression With Interaction Term

4. With the addition of an interaction term, the R^2 increases to .407. Is this statistically significant?
5. The interaction coefficient (the coefficient for c10, the interaction variable) is significant at the .05 level but not at the .01. (The obtained probability is .025.)

H. These values can be compared by using the formula described in Class 14 as an alternative test for significance of the interaction term.

$$F_{obs} = \left(\frac{\frac{(R^2_{complete} - R^2_{reduced})}{g}}{\frac{(1 - R^2_{complete})}{(N - K - 1)}} \right)$$

1. Note: I have simply rewritten the formula slightly.

2. As indicated before, N is the number of cases (here 144), K is the total number of variables in the complete model (here 3: 1 for Z, 1 for X, and 1 for W, the interaction term), and g is the number of variables left out of the "reduced" model (here 1).
   i. Remember the reduced model, the one without the interaction, is being compared to the complete model to see if the extra term(s) add anything significant.
   ii. The degrees of freedom for testing this model are **g** and **N - K - 1**.
3. For the data at hand the observed F is:

$$F_{obs} = \left(\frac{\frac{(.401 - .385)}{1}}{\frac{(1 - .407)}{(144 - 3 - 1)}} \right)$$

$$= 5.1939$$

4. The critical value of F at the .05 level with g = 1 and N - K - 1 = 140 degrees of freedom is 3.84; at the .01 level it is 6.83. Consequently we can reject at the .05 level (but not the .01 level) the null hypothesis that the partial β for interaction is zero.
   i. To be most complete we should report the obtained value of the probability of F under the null hypothesis:

$$.05 \quad prob(F = 5.1939 \quad H_0) \quad .01$$

5. We conclude that the relationship between Perot and Reagan voting is differs in Oregon from the other states considered.
6. Incidentally, as noted before, an F with 1 and N - K -1 degrees of freedom is the square of a t with N - K - 1 degrees of freedom. Consequently, when we take the square root of this F, we obtain $t_{obs} = 2.279$, the same value reported in Table 5 above.
   i. The moral is that when is looking at a single additional variable, one can use the t from the analysis based on the model in which it appears.

VI. NEXT TIME:

A. More on multiple regression.

1. Transformations
2. Plots of residuals

Go to Notes page

Go to Statistics page