DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

REGRESSION, INFERENCE AND CAUSATION

I. AGENDA:
- A. Simultaneous confidence intervals
- B. Multicolinearity
- C. Causal inference and experimental and quasi-experimental designs
  - 1. A useful procedure for making transformations
- D. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition, pages 541 to 543 and Chapter 10 (again).

II. SIMULTANEOUS CONFIDENCE INTERVALS:
- A. When dealing with two variables we have seen how to construct confidence intervals for individual parameters.
  - 1. The basic ideas: in addition to "point" estimators of β_0 and β_1, we want intervals such that with $100(1 - \alpha)$ percent confidence we can say that our estimation procedure includes the true values.
  - 2. $100(1 - \alpha)$% Intervals for β_1

$$\hat{\beta}_1 \pm t_{(\alpha,\, N\,-\,2)}\hat{\sigma}_{\hat{\beta}_1}$$

    - i. The upper and lower limit or bound have the property that if we take repeated samples of size N from a population in which the regression constant is β_1, 95 percent of our intervals will contain the true value.
  - 3. And we could construct a similar interval for the regression constant.
- B. Moreover we could find confidence intervals for partial regression coefficients in the same way.
  - 1. In fact doing so is standard procedure in most applications and published research.
- C. We might call this "one-at-a-time" intervals.[1]
  - 1. But actually we usually want to construct several intervals with the same sample data at the same time.
  - 2. We would thus want a level of confidence that applies **simultaneously** to

[1]Douglas C. Montgomery and Elizabeth A. Peck, ***Introduction to Linear Regression Analysis*** (1992) page 32

all of the intervals.

i. For example, if we have two estimators, and two sets of intervals that are constructed separately from each other but with the same data, the total probability that both intervals cover the parameters is not $(1 - \alpha)$ but rather $(1 - \alpha)^2$

ii. If we had two sets of 95 percent intervals, the probability that both contain their population values is $.95^2 = .9025$, not .95 as we might think.

iii. Furthermore since we are using the same set of data, these interval estimators are not independent of one another.

D. One solution is to construct a confidence region or ellipsoid, but we will use a simpler technique.

1. In particular, the confidence intervals will be

$$\hat{\beta}_k \pm \nu \, \hat{\sigma}_{\hat{\beta}_k}$$

i. Where ν is chosen so that the specified probability that **all** the intervals contain the true values is $1 - \alpha$, as we think.

1) As you might guess, ν is a critical value such as a t.

ii. Example, ν is chosen so that the probability the all intervals are correct is .95.

iii. Or equivalently, if we are estimating, say, K parameters and want simultaneous 95 percent intervals for them, we will choose ν so that the K set of intervals

$$\hat{\beta}_k - \nu \, \hat{\sigma}_{\hat{\beta}_k} \leq \beta k \leq \hat{\beta}_k + \nu \, \hat{\sigma}_{\hat{\beta}_k}$$

1) are correct with probability $1 - \alpha = .95$.

E. The form of the intervals is exactly the same as before except for ν, which has to be chosen to make the statements true.

1. Fortunately for our purposes the choice is easy.
2. The method is called the **Bonferroni** intervals.

F. Bonferroni intervals:

1. Suppose we have K independent variables.
   i. Hence the model has K + 1 parameters
2. Moreover suppose we want intervals for r of these parameters.
   i. Normally r would equal K + 1 or K.
   ii. In these circumstances we use for ν:

$$t_{\alpha/2r,\,(N - K + 1)}$$

iii. This is just a t with N - 2 degrees of freedom.

iv. The only thing different is that we adust alpha (the level of significance) by r, which is the number of intervals.

v. So instead of looking for the α in the t table, we look for $\alpha/2r$.

3. Examples:

i. If we had N cases and a simple one variable regression model with K + 1 = 1 + 1 = 2 parameters and wanted simultaneous intervals for both (r = 2).

1) For 95 percent intervals α = .05 and we need the .05/4 = .0125 level.

a) If N = 20, then the t would be (using 18 degrees of freedom) roughly 2.552.

2) For 99 percent simultaneous intervals we divide .01 by 4.

ii. If we had 6 independent variables, including possibly some for dummy variables and interaction terms, and want intervals for them but not the constant, then r = 6 and

1) for 95 percent intervals we would use .05/6 = .008.

a) t tables don't, of course, contain this level of significance, so we would probably use the .001 level.

b) If N = 27, then the degrees of freedom are 20 and the t would be very approximately 2.845.

c) If we wanted 99.9 percent intervals, then .001/6 = .0001 would be the required level of significance.

1) Again, just use the smallest value in the table.

G. Notes: this is a very conservative procedure in that we could accept some null hypotheses (that is, have too large intervals) more than we should.

1. But for most social science applications it should be as good a method as simply constructing one-at-a-time intervals.

H. Numerical examples

1. Here are the results for an air quality model that we found to be acceptable.

```
Mortality = 1155 + 0.252 SO2 - 24.8 Educat + 3.71 %Nonwht

59 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant      1154.99       72.15      16.01    0.000
SO2           0.25182     0.08390       3.00    0.004
Educat        -24.773       6.327      -3.92    0.000
%Nonwht        3.7123      0.5899       6.29    0.000

S = 39.26       R-Sq = 62.5%     R-Sq(adj) = 60.5%
```

2. Suppose we want to construct simultaneous confidence intervals for the three partial regression parameters r = 3 but not the constant.
   i. The degrees of freedom are 55.
   ii. Let's construct 95 percent intervals so $\alpha = .05$.
   iii. Moreover, .05/3 = .02 (about) so we'll use an appropriate two-sided t value from the table with "infinite" degrees of freedom.
      1) The Agresti and Finlay table stops with 29.
      2) The value is 2.326
   iv. Hence the intervals are:

$$\mathbf{.25182 + .08390(2.326)}$$
$$\mathbf{.055 - .195}$$

$$\mathbf{-24.773 \pm 6.327(2.326)}$$
$$\mathbf{-39.49 - -10.06}$$

$$\mathbf{3.7123 \pm .5899(2.326)}$$
$$\mathbf{2.34 - 5.084}$$

III. MULTICOLINEARITY:

A. As Agresti and Finlay point out (***Statistical Methods***, 3rd edition, page 541), independent variables--especially the ones social scientists study--often overlap in the sense that there are correlations among them.

1. In fact, many independent variables are highly inter-correlated as we have seen several time.s
   i. Consider the relationships among education, income, social status, housing prices and so forth that are used to explain, say, crime or voting turnout.
   ii. In some senses these variables measure the same things
   iii. Alternatively we might consider them **indicators** of some underlying concept such as social status.
   iv. For example, consider this figure:

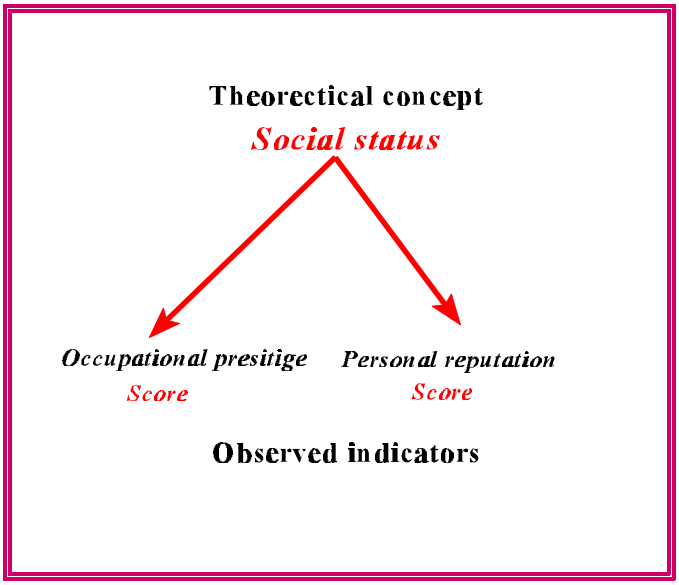

Figure 1: Latent and Observed Variables

v. An investigator might have two "paper-and-pencil" survey questions that measure occupational and personal prestige.

vi. Since they both measure "status," however, we would expect these indicators to be (highly) correlated.

vii. If we then use them in a regression model to explain a some dependent variable, we will be dealing with multicolinearity.

B. The situation arises frequently in policy and social research.

1. Here are some data from the famous Coleman[2] report on the effects of school desegregation.

i. The data consist of averages (means) from 20 schools

ii. The variables are:

1) Y (c1) mean verbal scores for sixth grade students.
2) X_1 (c2) Staff per pupil
3) X_2 (c3)Percent of students whose fathers are white collar
4) X_3 (c4) SES composite (means of family size, family intactness, fathers' education, percent white collar)
5) X_4 (c5)Mean teachers' verbal scores.
6) X_5 (c6) Mean mothers' education (1 unit = 2 years)

[2]J. S. Colman and others, ***Equality of Educational Opportunity***, 2 volumes, Office of Education, Department of Health, Education, and Welfare, 1966. Cited in Frederick Moststeller and John Tukey, ***Data Analysis and Regression*** (1977) page 556.

Verbal	Staff	White Collar	SES	Teacher Verbal	Mother's Education
3.83	28.87	7.20	26.60	6.19	37.01
2.89	20.10	-11.71	24.40	5.17	26.51
2.86	69.05	12.32	25.70	7.04	36.51
2.92	65.40	14.38	25.70	7.10	40.70
3.06	29.59	6.31	25.40	6.15	37.10
2.07	44.82	6.16	21.60	6.41	33.90
2.52	77.37	12.70	24.90	6.86	41.80
2.45	24.67	-0.17	25.01	5.78	33.40
3.13	65.01	9.85	26.60	6.51	41.01
2.44	9.99	-0.05	28.01	5.57	37.20
2.09	12.20	-12.86	23.51	5.62	23.30
2.52	22.55	0.92	23.60	5.34	35.20
2.22	14.30	4.77	24.51	5.80	34.90
2.67	31.79	-0.96	25.80	6.19	33.10
2.71	11.60	-16.04	25.20	5.62	22.70
3.14	68.47	10.62	25.01	6.94	39.70
3.54	42.64	2.66	25.01	6.33	31.80
2.52	16.70	-10.99	24.80	6.01	31.70
2.68	86.27	15.03	25.51	7.51	43.10
2.37	76.73	12.77	24.51	6.96	41.01

7) It should be clear that the independent variables, the ones used to explain students' verbal scores are highly correlated among themselves.

C. Consequences of colinearity among predictors.

1. Consider two predictors, X_1 and X_2 regression.
2. The standard deviation or error of the estimator of β_1 can be shown to be"

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{1}{1 - R^2_{X_1X_2}} \left[\frac{\hat{\sigma}_{Y|X_1X_2}}{\sqrt{\sum (X_1 - \bar{X}_1)^2}} \right]$$

i. Although the formula may look formidable, it is very similar to the ones we saw before.

ii. The main new factor is the multiple correlation between the X's.

iii. Look at what happens if this R^2 gets close to 1. Then the expression to the left also gets very large because we would be dividing 1 by a number that is almost 0. (If $R^2 = .9999$, for example, $1 - R^2$ is going to be very small and when we divide that number into 1, we obtain

a large number.)

iv. Consequently the expression on the right gets multiplied by a large number.

v. The long an the short is that the standard error of the estimator becomes too large, which affects in turn significance tests and confidence intervals.

vi. In particular, confidence intervals will be too wide and we will be accepting too many null hypotheses that β_k is zero.

3. On a more practical level, one of the consequences of independent variables that are themselves highly inter-correlated is that, although parameter estimators are unbiased, they are "unstable" in that their values from sample to sample "jump" around quite a bit.

i. Frequently the signs change just by adding or subtracting a variable from a model.

D. Example:

1. Although I recommend a more systematic approach, let's just regress verbal scores on all of the independent variables.

2. Here's the result:

```
The regression equation is
Verbal = 0.29 + 0.0062 Staff + 0.0425 White + 0.230 Ses - 0.182
T eachverb - 0.0731 Mothers

Predictor        Coef       StDev          T        P       VIF
Constant        0.292       2.956       0.10    0.923
Staff         0.00615     0.01089       0.57    0.581       8.6
White         0.04246     0.03349       1.27    0.226      11.3
Ses           0.22966     0.08658       2.65    0.019       1.4
Teachver      -0.1819      0.4184      -0.43    0.670       8.1
Mothers      -0.07311     0.05029      -1.45    0.168       9.3

S = 0.4190      R-Sq = 37.3%     R-Sq(adj) = 14.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5      1.4604      0.2921      1.66    0.208
Residual Error    14      2.4575      0.1755
Total             19      3.9179
```

3. At first (and quick) glance the data seem to fit a linear model.

4. But closer inspection shows that besides most of the partial coefficients not being significant some have signs counter to what one might expect.

5. The ones relating staff and mothers' education are "wrong," which means

that the signs seem conuter-intuitive.

6. When we regress scores on just staff, the sign changes from negative to positive.
7. The coefficient is: $\hat{\beta}_{Y|X_1\ldots} = .17$

E. A somewhat useful tool for checking on the effects of multicolinearity is the so-called **variance inflation factor** (**VIF**), which can be defined as follows:

$$VIF_k = \left(\frac{1}{(1 - R_k^2)} \right)$$

i. where R^2_k is the multiple R obtained from regressing the X_k (the kth independent variable) on all the other predictor or independent variables.

1. MINITAB calculates the VIF for each independent variable if you check the option.
   1) Look in the options box.
2. An inflation factor greater than 10 or even 5 implies that some of the predictor or regressor variables are highly related and may be creating instability in the coefficients. By "instability" I mean that if you leave out a variable, the significance of included variables changes dramatically or the numerical value of the coefficients is affected by including or excluding a few case. Bluntly stated, if your model building doesn't seem to make sense, it's likely that some of the independent variables are highly inter-related.
   i. Unfortunately, the use of interaction variables, which involves multiplying one factor by another and thus creating a dependence among variables, can create exacerbate the problem.

F. What to do?
1. Always obtain a matrix of correlations among independent variables.
2. Look for strong interrelationships
   i. How “strong” is hard to specify at this point.
3. As indicated previously consider dropping redundant variables or combining them in a composite measure.

IV. CAUSAL INFERENCE AND NON-EXPERIMENTAL RESEARCH:

A. Here are some hypothetical data. Consider two "treatments" for an illness (or welfare or criminality or whatever) The "success" rates are reported as follows:

TREATMENT I	TREATMENT II
60%	40%

B. What inferences can one draw from these data, often called "crude success rates"?
   1. This sort of problem, which explains why statisticians and clinicians are so fussy about research design, comes up again and again in the social and policy sciences.
   2. The answer, as we will see later, is not too much. Certainly, we cannot conclude on the basis of these data that the first treatment is "better" than the second. In fact, unless we know more, it could easily be the case that the second is much more efficacious.
   3. Most important, we should until we know more avoid making any sort of causal inference

C. Demonstrating causality: the "traditional" social science view:
   1. Constant conjunction (covariation)
   2. Temporal order (an effect cannot be its cause)
   3. Elimination of alternative hypotheses or explanations

V. EXPERIMENTAL DESIGN:

A. Consider this quotation from the Attorney General's "Final Report" on pornography: "In both statistical and experimental settings exposure to sexually violent materials has indicated and increase in the likelihood of aggression."
   1. What is being asserted is that "exposure" to "sexually violent" movies and magazines is not only associated with aggression but a <u>cause</u> of it.
   2. Here are some (<u>hypothetical</u>) data that might pertain to this issue:

		MEN WHO WATCH PORNOGRAPHIC (X-RATED) MOVIES:			
		Never	1/Month	1/Week	2/Week
Report having violent fantasies about women:	Yes	25%	35%	45%	60%
	No	75	65	55	40
Totals:		100% (50)	100% (75)	100% (40)	100% (35)

Table 1: Hypothetical Experimental Data

3. Two questions immediately arise:
   i. Are the variables--"Viewing X-rated movies" and "Occurrence or non-occurrence of violent fantasies" associated or related?
   ii. Is viewing a causal factor that produces aggressive fantasies and possible aggression itself?
4. The problem is this: true, men who see a lot of pornography seem to have more aggressive thoughts than men who do not see such movies (compare the percents). But, and this is the 64-thousand dollar question, do the movies cause the fantasies or do men with such fantasies already in their minds go to these movies (perhaps to gratify them) while men without such images do not. In other words, it is possible that self-selection is operating.
5. One can picture these alternatives with the use of causal diagrams in which arrows indicate direct causation and the absence of arrows the lack of direct causation.

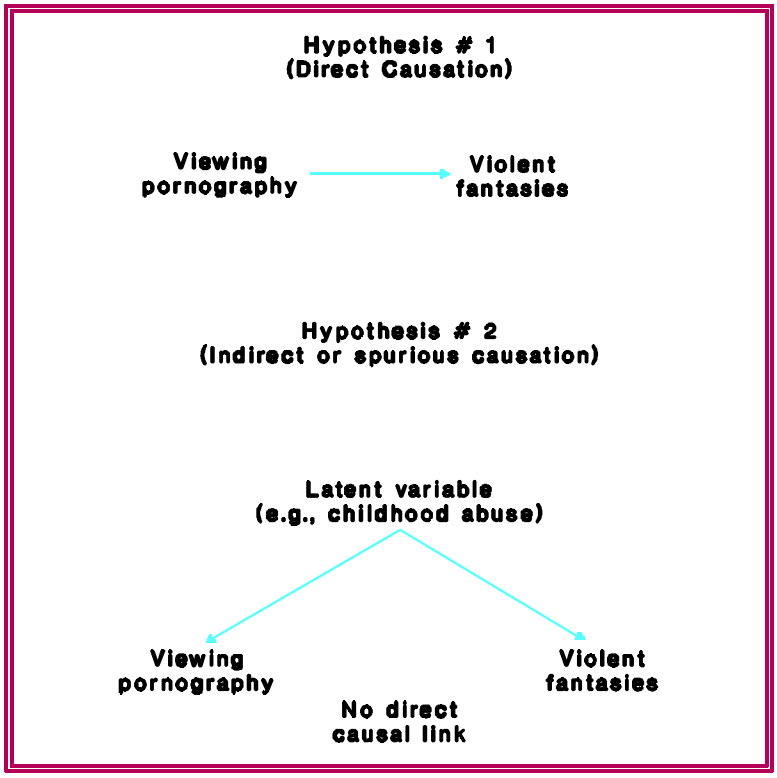

Figure 2: Causal Models of The Effects of Pornography

VI. THE "CLASSICAL" RANDOMIZED EXPERIMENT:

A. Causal inferences, it is often argued, can be made in the context of a randomized, controlled experiment.

1. Even though social scientists frequently cannot experiment on subjects, the logic of the procedure is worth investigating.

B. It (the simplest possible version) has this general form

		Time 1		Time 2
R	Experi-mental group	Y_{Pre}	X	Y_{Post}
R	Control group	Y'_{Post}		Y'_{post}

C. Notation:

1. **R** stands for randomization: the random assignment of cases to the experimental and control group.
   i. Randomization is what really gives an investigator control.
2. X is the experimental "maniupulation."

D. Assumption (because of randomization):

$$Y_{pre} = Y'_{pre}$$

1. The Y's are measures of the dependent variable, number of violent fantasies in our example.

E. Measurement of effects:

1. Main effect:
   i. What we expect to find if the "treatment" or experimental manipulation had an effect:

$$\Delta = Y_{post} - Y'_{post}$$

and

$$\Delta \neq 0$$

2. Moreover we would also expect:

$$\Delta' = Y'_{post} - Y'_{pre} = 0,$$

except for sampling error

F. Internal validity
   1. Definition: did the experimental variable in fact make a difference in this instance?
   2. Internal validity is the fundamental question: was there anything about the procedure that could have produced a large **Δ**, aside from X, the supposedly causal variable.
   3. Without internal validity one cannot accept the causal attribution.
   4. Factors affecting internal validity
      i. History: specific events occurring between first and last measurements of Y. Example: appearance of newspaper report on pornography during an experiment.
      ii. Testing: the effects of being measured, of being asked about "pornography," of being told one is in an experiment, etc.
      iii. Maturation of subjects: the participants change as the experiment proceeds. Respondents are sensitized, for example.
      iv. "Demand characteristics": the subjects anticipate and act out the experimenter's objectives. ("I've seen a violent X-rated movie, so I should act violently because that's what the investigator wants." This will seldom be a conscious decision but it may be part of people's motivation anyway.)
   5. Research design tries to minimize these problems.

G. External validity:
   1. Definition: To what populations can the results be generalized? is anything about the subjects, the experimental setting, the measures, etc. that might be "unrealistic." Would you, for example, on the basis of the experiments you read support a restriction on the distribution of X-rated movies because the research shows them to be "potentially harmful?"

H. Some Fallacious Designs:
   1. Here are some faulty experimental designs--faulty in the sense that they do not allow one to unambiguously make causal inferences.
   2. No control Group:

	Time 1		Time 2
Exp Group	Y_{Pre}	X	Y_{Post}

   3. Since there is no control group, one cannot say for sure that **Δ** is due to the X factor. What if the subjects would have changed anyway? A control group is almost always necessary.

i. Examples: enforcement of drunk driving enforcement, helmet laws, 55 mph speed laws

4. One-shot study:

X	Y_{Post}

i. Here there is no comparison at all so no causal inferences seem warranted.

I. Many traditional explanations seem to fall into this category.

1. No randomization: simple comparison

VII. TIME SERIES:

A. See the figure below. The idea of time series analysis is that a variable, the rate of crime or out-of-wedlock births, for example is increasing or decreasing over time.

1. The idea is that a factor or condition or policy represented by X "causes the trend (the increase or decrease) to decline.

2. The problem is how do we know that X, and not W or Z, is responsible for the change? It is difficult to make causal inferences in non=experimental time series analyses.

$...Y_{t-3}, Y_{t-2}, Y_{t-1}, Y_t, X\ Y_{t+1}, Y_{t+2}, Y_{t+3}...$

Y's are measurements on the dependent at different times.
X is the "experimental" variable; that is, the event or "intervention" that supposedly "caused" a change in the time series.

3. We start analyzing time series soon.

VIII. NEXT TIME:

A. Additional material on regression.

Go to Notes page

Go to Statistics page