Psychology Seminar
Psych 406
Dr. Jeffrey Leitzel

Structural Equation Modeling
Topic 1: Correlation / Linear Regression

Outline/Overview

- Correlations (r, pr, sr)
- Linear regression
- Multiple regression
  - interpreting b coefficients
  - ANOVA model test
  - R^2
  - Diagnostics
    - Distance(Discrepancy), Leverage, Influence, & Multicollinearity

The correlation coefficient (Pearson r)

- Continuous IV & DV (Interval/Ratio level data)
- Dichotomous variables
- Strength of linear relationship between variables
- r^2 coefficient of determination

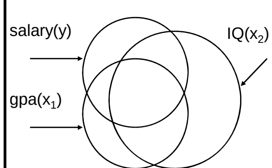
$$r = \frac{\Sigma Z_x Z_y}{N - 1}$$

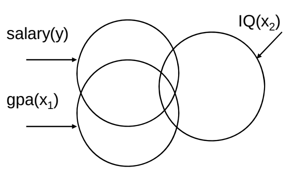
Partial and semipartial correlations

- Partial Correlation
  - r between two variables when one or more other variables is partialled out of both X and Y
  - example: experimenter interested in investigating the relationship between income and success in college
    - measured both variables, ran correlation
    - it was statistically significant
- he began repeating these results to all of his students
- do well in college = large salaries
- what might the bright student in the back of one of his classes have raised as an issue?

Partial correlation (cont.)

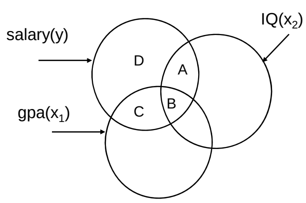
- IQ as a third variable effecting both college success and salary earned
- The partial correlation between college success and salary with IQ partialled out of both variables





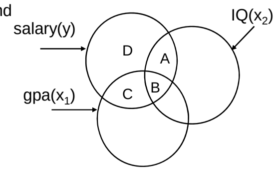
Partial correlation (cont.)

- correlation between the residuals is the partial correlation between income and success, partialling out IQ
- generally represented by $r_{y1.23...p}$
  - subscripts to the left of the dot represent variables being correlated
  - those to the right of the dot are those being partialled out
- $r_{y1.2}$=correlation between salary and GPA with IQ partialled out
- $pr^2=c/(c + d)$



Semipartial correlation

- also called the part correlation, far more routinely used
- $r_{y(1.2)}$ represents the correlation between the criterion (y) and a partialled predictor variable
- the partial correlation has variable x_2 partialled out of both y and x_1
- for the semipartial, x_2 is partialled only out of x_1
- the correlation between y and the residuals from x_1 predicted by x_2
- correlation between y and the part of x_1 that is independent of x_2
- $r_{y(1.2)}$=correlation between salary and GPA with IQ partialled out
- $sr^2=c/(a + b + c + d)$



Semipartial correlation (cont.)

- can rearrange semipartial correlation formula:
  - $R^2_{y.12} = r^2_{y2} + r^2_{y(1.2)}$
- R (Multiple correlation) based on information from variable 1 plus additional <u>nonredundant information</u> from variable 2 through variable p so: $R^2_{y.123...p} = r^2_{y1} + r^2_{y(2.1)} + r^2_{y(3.12)} +...+ r^2_{y(p.123...p-1)}$
- If the predictors were completely independent of one another, there would be no shared variance to partial out
- would simply sum the r^2 for each X variable with Y to get the Multiple R^2

Simple linear regression

- Finding the equation for the line of best fit
- Least squares property
  - Minimizes errors of estimation
- The line will have the form: $y' = a + bx$

  Where: y' = predicted value of y
  a = intercept of the line
  b = slope of the line
  x = score of x we are using to predict y
- Multiple regression an extension of this

Multiple regression models

- Simple linear regression models generally represent an oversimplification of reality
- Usually unreasonable to think in terms of a single cause for any variable of interest
- Our equation for simple linear regression: $y' = a + bx$
- becomes: $y' = a+b_1x_1+b_2x_2+\ldots+b_kx_k$

Multiple regression (cont.)

- $y'=a+b_1x_1+b_2x_2+\ldots+b_kx_k$
- Where y = dependent variable (what we are predicting)
- $x_1, x_2,\ldots, x_k$ = predictors (independent variables or regressors)
- $b_1, b_2,\ldots, b_k$ = regression coefficients associated with the k predictors
- a = y intercept
- The predictors (x's) may represent interaction terms, cross products, powers, logs or other functions that provide predictive power

Interpreting b coefficients

- If x_1 is held constant, every unit change in x_2 results in a b_2 unit change in y'
- The constant, a (y intercept) does not necessarily have a meaningful interpretation
  - any time 0 is outside the reasonable range for predictors the intercept will be essentially meaningless

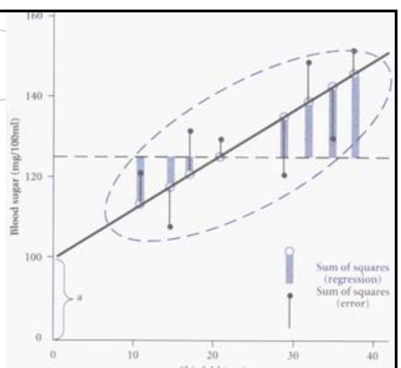
Overall ANOVA

- Testing overall model adequacy
- Null hypothesis H_0: $b_1=b_2=\ldots=b_k=0$
  Alternative hypoth H_1: any b <> 0
- F statistic (ANOVA) = $MS_{regression}/MS_{error}$
  $MS_{regression} = SS_{regression} / k$
  $MS_{error} = SS_{error} / n-(k+1)$
- Where n = number of observations
  k = number of parameters in model (excluding a)

ANOVA (cont.)

- $SS_{regression} = \Sigma(y' - M_y)^2$
- $SS_{error} = \Sigma(y - y')^2$
- When the overall regression model provides no predictive power the two mean square quantities will be about equal
- We are testing an F statistic with (k, n-(k+1)) degrees of freedom

ANOVA in multiple regression



Output from Regression Analysis

Variable	Coefficient	SE	t	Significance
Constant	98.00	12.3	7.97	0.0001
Skinfold	1.14	0.046	24.70	0.0001

R^2 and Adjusted R^2

- Multiple R^2, coefficient of determination
- Total proportion of variance accounted for by all predictors in the model
- High R^2, values do not necessarily mean that a model will be useful for predicting Y
- Overfitting (adding essentially irrelevant predictors) can result in a high R^2 in the absence of any predictive power
- If we fit a regression model with n-1 predictors R will always be = 1.0 unless we have a rare data set where two cases are identical on all predictors but differ on Y value
- Adjusted R^2 takes into account our n and the number of predictors we have used in the model

Misuses/problems

- Multicollinearity
  - exists when two or more independent variables (predictors) contribute redundant information to the model
  - can think of conceptually as a problem in assigning unique variance components to variable
  - creates computational problems and coefficients that do not make sense
  - resolve by removing redundant predictor(s)
- Predicting outside the range of data that was used to generate the regression model
  - positive linear relationship
  - relationship may change-resulting in substantial error of estimation
- Failure to explore alternative models
  - need to be familiar with data
  - examine relationships between variables
  - be aware of potential interactions and include in model if necessary
  - try different models

Standardized betas

- with more than one independent variable:
  - b coefficients cannot be directly compared due to different scales across measures
  - standardized Beta coefficients allow us to compare the relative predictive power of variables in the equation
  - we get standardized Betas by converting all of our predictors to z-scores and running the regression

Regression diagnostics

- help us assess the validity of our conclusions and their accuracy
- careful screening of the data cannot be overemphasized
- Outliers
  - outliers: data points that lie outside the general linear pattern (midline is the regression line)
  - removal of outliers can dramatically affect the performance of a regression model
  - should be removed if there is reason to believe that other variables not in the model explain why the cases are unusual these cases may need a separate model

Outliers (cont.)

- outliers may also suggest that additional explanatory variables need to be brought into the model
- multivariate outliers can be difficult if not impossible to identify by visual inspection
- ex. 115 lbs. and 6 feet tall
  - either observation alone is not unusual, but together they are.
- unusual combinations might include odd pairings of scores on separate measures of ability or different indices of production
- multivariate outliers represent results that when considered as a whole do not make sense

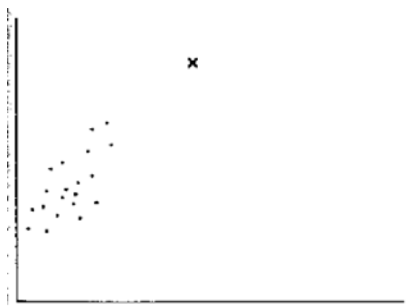
Distance, leverage, & influence

- most common measure of distance (also called discrepancy) is the simple residual
- distance between any point and the regression surface
- identifies outliers on the dependent (y) variable
- leverage statistic, h, also called *hat-value*, identifies cases which may influence the regression model more than others
- with reasonably large n ranges from essentially 0 (no influence on the model, actually the min is 1/n) to 1

Distance, leverage, & influence (cont.)

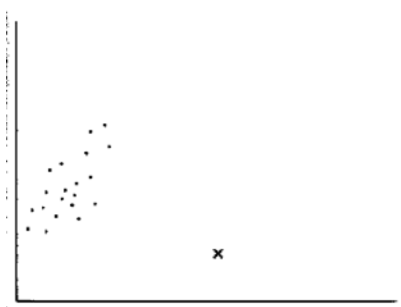
- leverage identifies outliers in the dv's
- <.2 fine, >.5 case should be examined
- cases that are high on distance or leverage can strongly influence the regression but do not necessarily do so
- points that are relatively high on both distance and leverage are very likely to have strong influence

Influence on regression line



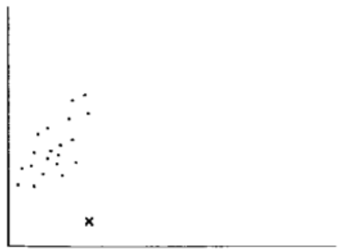
(a) High leverage, low discrepancy, moderate influence

Influence on regression line



(b) High leverage, high discrepancy, high influence

Influence on regression line



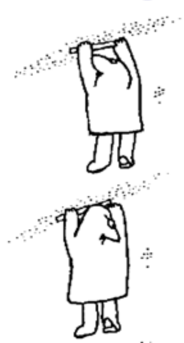
(c) Low leverage, high discrepancy, moderate influence

FIGURE 5.1 THE RELATIONSHIPS AMONG LEVERAGE, DISCREPANCY, AND INFLUENCE.

Cook's distance (D)

- most common measure of influence of a case
- it is a function of the squared change that would occur in y' if the observation were removed from the data
- interested in finding cases with D values much larger than the rest
- cut-off influential cases, D greater than 4/(n - k - 1)
- where n is the number of cases and k is the number of independents
- Cook's distance obtained by issuing "postestimation" command in Stata

Regression diagnostics



Diagnostics put to use

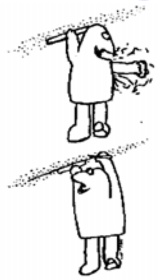


Figure 10.1. Regression diagnostics in action.

SOURCE: Reprinted with permission from the announcement of the Summer Program of the Inter-University Consortium for Political and Social Research, 1990.

Multicollinearity

- intercorrelation of independent variables
- R^2s near 1 violate assumption of no perfect collinearity
- high R^2s increase the standard error of the beta coefficients and make assessment of the unique role of each independent difficult or impossible
- simple correlations tell something about multicollinearity
- preferred method of assessing multicollinearity is to regress each independent on all the other independent variables in the equation

Multicollinearity (cont.)

- inspection of the correlation matrix reveals only bivariate multicollinearity, for bivariate correlations > 0.90
- to assess multivariate multicollinearity, one uses tolerance or VIF
- may not be any extremely high bivariate correlations
- if any variable can be represented as a linear combination of other variables in the model, perfect multicollinearity exists

Tolerance/Variance Inflation Factor (VIF)

- Tolerance
  - 1 - R^2 for the regression of that independent variable on all other independents, ignoring the dependent
  - as many tolerance coefficients as there are independents
  - higher intercorrelation of the independents, the tolerance will approach zero
  - part of the denominator for calculating the confidence limits on the b (partial regression) coefficient
- Variance inflation factor (VIF)
  - reciprocal of tolerance
  - when VIF is high there is high multicollinearity and instability of the b coefficients

Variance inflation factor (VIF)

R_j	Tolerance	VIF	Impact on SE_b
0.0	1.00	1.00	1.0
0.4	0.84	1.19	1.09
0.6	0.64	1.56	1.25
0.75	0.44	2.25	1.5
0.8	0.36	2.78	1.67
0.87	0.25	4.00	2.0