

# Multiple Regression

## Advanced Issues

# Correlation

- Any multivariate technique concerns the relationships among variables
- Correlation is a measure of the strength and direction of a *linear* relationship between two variables
  - Two variables' values vary about their mean (variance)
  - They may also covary with one another, having the tendency to move in the same or opposite directions (covariance)
- As such a correlation is an effect size in and of itself<sup>1</sup>
- When squared it represents the shared variance between the two variables

# Factors affecting correlation

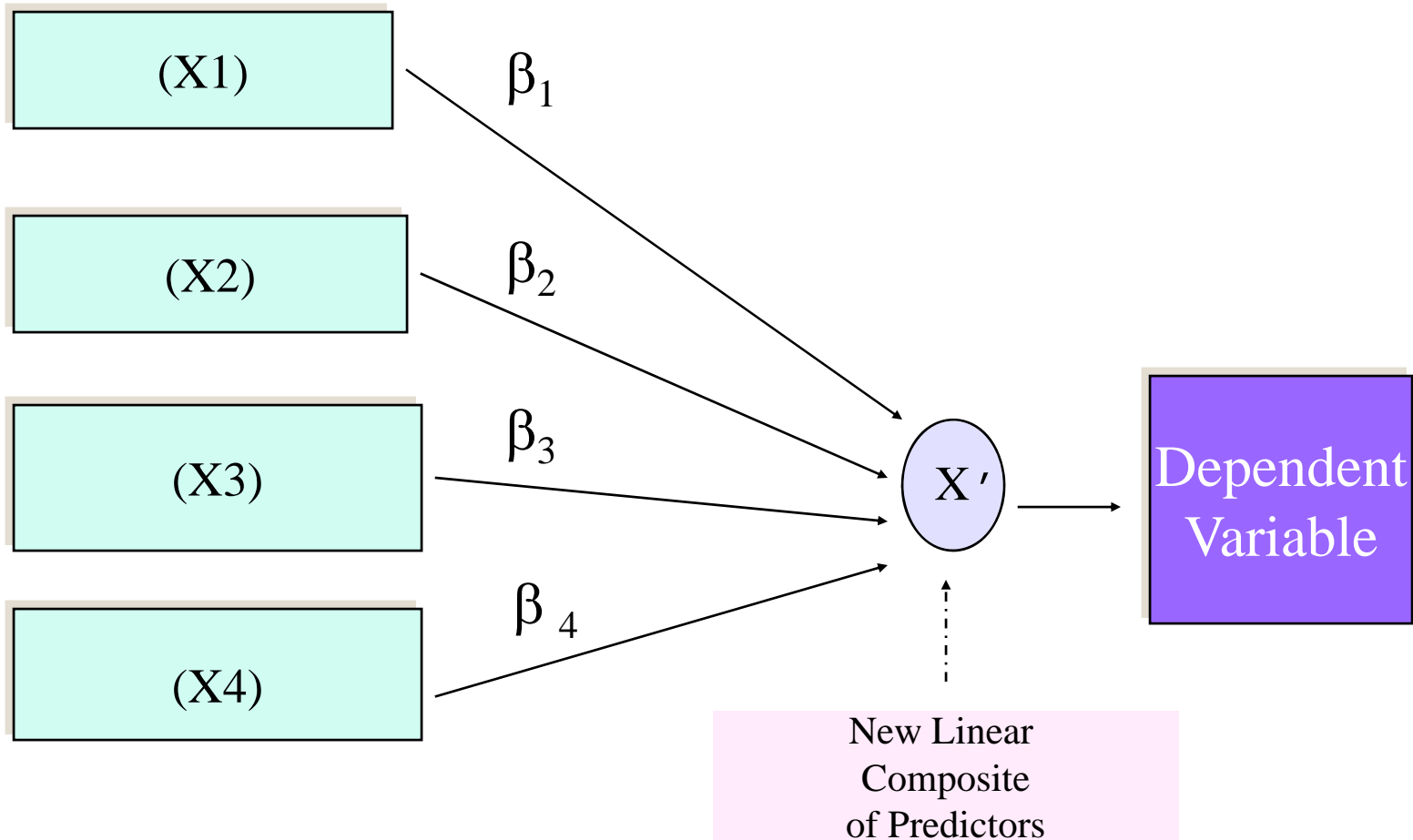
- **Linearity**
  - Nonlinear relationships will have an adverse effect on a measure designed to find a linear relationship
- **Reliability of measurement**
  - If you are not reliably distinguishing individuals on some measure, you will not capture the covariance that measure may have with another accurately
- **Heterogeneous subsamples**
  - Sub-samples may artificially increase or decrease overall  $r$ .
  - Solution - calculate  $r$  separately for sub-samples & overall, look for differences
  - Can be caused by lack of reliability
- **Range restrictions**
  - Limiting the variability of your data can in turn limit the possibility for covariability between two variables, thus attenuating  $r$ .
  - Common example occurs with Likert scales
    - E.g. 1 - 4 vs. 1 - 9
  - However it is also the case that restricting the range can actually increase  $r$  if by doing so, highly influential data points would be kept out
    - Wilcox 2001
- **Outliers can artificially increase or decrease  $r$**

# Regression

- Using the covariances among a set of variables, regression<sup>1</sup> is a technique that allows us to predict an outcome based on information provided by one or more other variables
- We also can get a sense of the overall correlation between a set of variables and some outcome
  - **Multiple R**
- Squaring that value has the same interpretation as before
- Regression equation:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

# Conceptual Understanding



# Interpreting Regression Results: Basics

- Intercept
  - Value of Y if X(s) is 0
  - Often not meaningful, particularly if it's practically impossible to have an X of 0 (e.g. weight)
- Slope, the regression coefficient
  - Amount of change in Y seen with 1 unit change in X
  - In MR that is with the others *held constant*
    - Standardized regression coefficient
      - Amount of change in Y seen in standard deviation units with 1 standard deviation unit change in X
      - In simple regression it is equivalent to the r for the two variables
- Standard error of estimate
  - Gives a measure of the accuracy of prediction
  - Model fit in terms of prediction
- $R^2$ 
  - Proportion of variance in the outcome explained by the model
  - Effect size
  - Model fit in terms of explanation

# Steps in Regression

- Determine whether prediction or explanation is the primary goal but maintain a balance between the two
- Specify the model
- Obtain good (not 'adequate'<sup>1</sup>) measures of the variables of interest
- Inspect the data, especially visually. Know the basics of the variables involved (descriptive information).
  - However, for MR this is not where you'll test model assumptions
- Run the initial analysis to test assumptions
  - Normality: bias, incorrect probability coverage
  - Homoscedasticity: model misspecification, inflated standard errors (inefficient estimates)
  - Independence: theoretical conclusions, inflated standard errors
  - Linearity: model misspecification
  - Lack of outliers: biased, inflated standard errors
  - Collinearity: inflated standard errors
- Graphical inspection
  - Basic scatterplots, Density/QQ plots of residuals, residuals vs. fitted, influence plots, component-residual plots etc.
- Statistical inspection
  - Any normality test on the residuals
  - Breusch-Pagan test for heteroscedasticity
  - Durbin-Watson for autocorrelation<sup>2</sup>
  - RESET test for linearity
  - Many measures of outliers (Cook's distance, dfBetas, Mahalanobis' distance etc.)
  - Variance inflation factor for collinearity

# Steps in Regression

- Do something about the problems
  - Compare initial output to a robust regression
  - Use bootstrapped estimates
  - Create composites or drop variables with collinearity
  - Run more appropriate models
- Validate the model

# Prediction vs. Explanation

- Doing a real regression involves much more than simply obtaining regression output from a statistical program
- First one must discern the goals of the regression, and there are two main kinds that are for the most part exclusive to one another in intent but do not have to be in practice
- Prediction
  - With prediction there is much less if any concern over variable importance etc.
  - One is more concerned with whether it (the model) works rather than how it works
    - Coefficients will be used to predict future data
  - Analytical example: Stepwise techniques
  - Applied example: prediction of graduate school success with incoming applicants
- Explanation
  - The goal here is to understand the individual relationships of variables with some outcome
  - Causal notions are implicit in this approach
  - Majority of psych usage falls almost entirely to this end
  - Analytical example: Sequential regression
  - Applied example: Personality factors involved in Aggression
- Putting them together: model averaging with validation and appropriate testing of variable importance

# Prediction vs. Explanation

- There is a problem with going to far on either end of the spectrum
- While understandable in some, if not many research cases, putting all emphasis on prediction without understanding well the ‘why and which’ would be untenable for most psych research
- However, if one is entirely on the explanation side, one often finds a ‘much ado about nothing’ type of scenario
  - No point in saying which predictor is more important with a crappy model
- Modern approaches allow one to establish predictive validity much more easily, as well as provide better means to understand variable importance

Research Goals

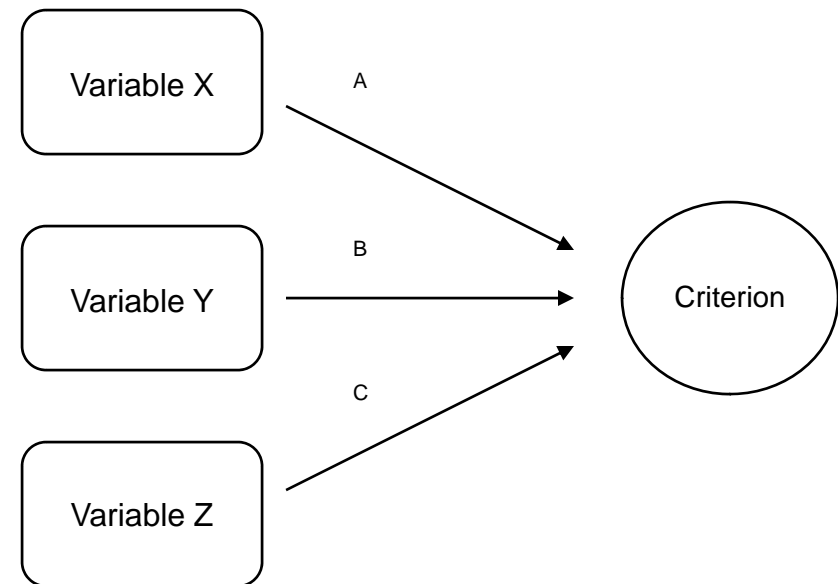
---

Explanation

Prediction

# Starting out: Model specification

- Specifying the model is obviously something to be done before data collection<sup>1</sup> unless one is doing a truly exploratory endeavor
- However one must consider relations (potentially causal ones) among *all* the variables under consideration
  - Given 3 variables of which none are initially set as an outcome, there are 20+ models which might best represent the state of affairs
- In addition one must consider interactions (moderating effects) among the variables
- Reliable measures must be used, or it will be all for naught
  - For some reason people try anyway



# Collect reliable data

- Reliability deals with measurement error (as opposed to sampling error or prediction error)
- In a simple bivariate setting, less reliable measures underestimate the true relationship between the variables
- In multiple regression, the relationships observed may vary wildly from the true relationship depending on the specifics
- In short, get good measures or suffer the consequences

# IED

- Initial examination of data includes:
- Getting basic univariate descriptive statistics for all variables of interest
- Graphical depiction of their distribution
- Assessment of simple bivariate correlations
- Inspection for missing data and determining how to deal with it<sup>1</sup>
- Start noting potential outliers, but realize that univariate outliers may not be model outliers
- The gist is you should know the data inside and out before the analysis is conducted<sup>2</sup>

# Running the Analysis

- ANOVA summary table- you should know every aspect of an ANOVA table and how it is derived
- In the following the Model df is the number of predictors  $k$ , the Error df is  $N - K - 1$
- The  $F$  is a ratio of variance attributable to the model and that which is unaccounted for
- The square root of the  $MS_{\text{error}}$  is the standard error of estimate (aka residual standard error), a measure of predictive fit
- $R^2$  is the  $SS_{\text{Model}} / SS_{\text{Total}}$ , the amount of variance in the DV explained by the model, a measure of explanatory fit
- Often times you will simply see in-text reporting of  $F(df_1, df_2) = \text{value}$ ,  $MSE = \text{value}$ ; As you can see though, that is enough information to reconstruct the entire ANOVA table.

Source	SS	df	MS	F	$R^2$
Model	$\sum (\hat{Y} - \bar{Y})^2$	$k$	$SS_{\text{Model}}/df_{\text{Model}}$	$MS_{\text{Model}}/MS_{\text{Residual}}$	$SS_{\text{Model}}/SS_{\text{Total}}$
Residual	$\sum (\hat{Y} - Y)^2$	$N - k - 1$	$SS_{\text{Res}}/df_{\text{Residual}}$		
Total	$\sum (Y - \bar{Y})^2$	$N - 1$			

# Statistical significance of the model

- Among other things the ANOVA summary table tells us whether our model is statistically adequate
  - $R^2$  different from zero
  - The regression equation is a better predictor than the mean
- As with simple regression, the analysis involves the ratio of variance predicted to residual variance
- As we can see, it is reflective of the relationship of the predictors to the DV ( $R^2$ ), the number of predictors in the model, and sample size
- Example ANOVA tables from SPSS and R from two different regressions<sup>1</sup>

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6.13E+10	3	2.042E+10	125.176	.000 <sup>a</sup>
	Residual	7.67E+10	470	163112654.7		
	Total	1.38E+11	473			

a. Predictors: (Constant), Previous Experience (months), Months since Hire, Educational Level (years)

b. Dependent Variable: Current Salary

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
speed	1	21185.46	21185.46	89.57	0.0000
Residuals	48	11353.52	236.53		

DV = vehicle stopping distance

# Back up

- Wait a second what just happened?
- You 'did' an MR analysis, but *how* did you do it?
- Click the icon
- You don't have to get crazy with it, but you should get a sense of how the data is manipulated to produce coefficients, regular or standardized
- You also need to be able to produce the model formula and predicted value given the coefficients<sup>1</sup>



# Focus on Prediction: Residuals

- The residual standard error tells you how far off your predicted values are from the original on ‘average’, where average in this case is like standard deviation is an average
  - Your average residual is zero, the RSE is the square root of the average squared residuals (i.e. the MSE in the ANOVA)
- This gives an initial assessment of model fit

# Focus on Prediction: Fit

- Other fit indices are geared toward this estimate but also take into account model complexity, penalizing models with more predictors
  - **Ockham's razor**
- For example, the Akaike (AIC) and Bayesian<sup>1</sup> (BIC) Information Criteria are good for relative model comparison
  - **The operative word being *relative*, i.e. you need more than one model to test. They are not easily interpreted by themselves. The better model is the one with the smaller value i.e., the more negative.**
- Other model fit indices will be seen with other modeling techniques
  - **SEM and path analysis: RMSEA, CFI, etc.**

TABLE 6  
Grades of Evidence Corresponding to Values of the Bayes Factor for  $M_2$   
Against  $M_1$ , the BIC Difference and the Posterior Probability of  $M_2$

BIC Difference	Bayes Factor	$p(M_2 D)(\%)$	Evidence
0-2	1-3	50-75	Weak
2-6	3-20	75-95	Positive
6-10	20-150	95-99	Strong
>10	>150	>99	Very strong

TABLE 8  
Minimum Percent Reduction in the Residual Sum of Squares Required for Dif-  
ferent Grades of Evidence in Favor of One Additional Variable in Linear Re-  
gression. When  $R^2$  is small, this is roughly equal to the required increase in  $R^2$ .

Evidence	Minimum BIC Difference	$n$					
		30	50	100	1,000	10,000	100,000
Weak	0	10.7	7.5	4.5	0.7	.09	.012
Positive	2	16.5	11.2	6.4	0.9	.11	.014
Strong	6	26.9	18.0	10.1	1.3	.15	.018
Very Strong	10	36.0	24.3	13.6	1.7	.19	.022

# Focus on Prediction: Coefficients and Validation

- The actual value of the coefficients is of course primary in determining predictive accuracy
- This in turn implies that they will be tested on new data, whether via replication, simple cross-validation by means of some form of sample splitting, or the bootstrap



# Focus on Explanation: Variable Importance

- As the goal of much of social science research is explanation, determining variable importance is a key concern
- Unfortunately, the way it is typically done is crude, difficult to make sense of, and not done with much thought
- Raw coefficients: always the first stop
- Standardized metrics
  - Standardized coefficient
  - Partial correlation
  - Semi-partial correlation
  - Average semi-partial correlation

# “Controlling for”

- Let's start with raw coefficients- What do they tell us?
- How much does Y change with a one unit change in X... *controlling for* the other variables in the model
- Now for some fun go ask random researchers what ‘controlling for’ actually means
- At a conceptual level it simply means we are considering this variable's effects on the DV with respect to the other predictors
- What it actually means is that this is the average effect (slope) seen at each of the levels of the other variables
  - E.g. average effect of sex role identity on marital satisfaction for each value of age
- It does not mean experimental control, and it also doesn't mean we've controlled for what might be potentially important variables that weren't included in the model

# Variable importance: Raw coefficients

- Raw coefficients tell you whether the predictor is important in and of itself, but with the standard error we can also determine statistical significance
  - If you cannot make this 'effect size' determination with your own research, you don't know your variables well enough to be doing the analysis in the first place
  - Standard error
    - measure of the variability that would be found among the different slopes estimated from other samples drawn from the same population
- Statistical significance however is typically not a good way to determine variable importance, and certainly not relative importance, and should probably only be noted casually if at all<sup>1</sup>
  - Arbitrary changes in coefficients can lead to notable differences in observed p-values
- However the standard error does allow us to get confidence intervals for the coefficients, which are desirable and should be a standard part of reporting

$$s_{y.12}^2 = \frac{SS_{res}}{N - k - 1}$$
$$s_{b1} = \sqrt{\frac{s_{y.12}^2}{\sum x_1^2 (1 - r_{12}^2)}}$$
$$t_{b1} = \frac{b_1}{s_{b1}}$$

# The standardized coefficient

- If we standardized our variables before running the regression analysis (i.e. used the correlation matrices) we would have a standardized regression coefficient
  - How much Y would typically change given a one standard deviation change in X
  - In the simple setting it equals the Pearson r, in MR it is a type of partial correlation
- We like this because most of the measures used in psychology are on arbitrary scales (e.g. Likert)
- In this manner we can more easily compare one variable to another in absolute terms
- However, it is not a justifiable method of determining one variable is more important than another just because their coefficients differ
  - Did you really think they'd be the same going into the analysis?
  - Did you decide beforehand what would be a meaningful difference?

# Other Standard Metrics

- Partial correlation
  - Predictor-DV correlation after partialling out the shared variance that both have with the other variables
  - Computationally:
    - $SS_{\text{predictor}} / (SS_{\text{predictor}} + SS_{\text{residual}})$
- Semi-partial Correlation
  - Predictor-DV correlation after partialling out the shared variance the DV has with other variables
  - When squared it represents the amount  $R^2$  would increase if it was added last to the model
  - Computationally:
    - $SS_{\text{predictor}} / SS_{\text{total}}$
- Average semi-partial correlation
  - Calculate the semi-partial correlation over all possible predictor models and average
  - Benefits: decomposes  $R^2$  into the relative contributions of the predictors (i.e. it adds up to the model  $R^2$  unlike the others)

# Problem

- None of these metrics are useful in variable importance without replication, validation or statistical tests (interval estimation)
- They are subject to sampling variability and the order of variable importance you happen to see with this sample is unlikely to hold unless differences are extreme and the dataset is very large
- In other words, claims cannot be made regarding variable importance without taking further steps
  - Validation is extremely important for both model prediction and explanation

## Focus on Explanation:

### The Multiple Correlation Coefficient and $R^2$

- The multiple correlation coefficient (Multiple R) is the correlation between the DV and the linear combination of predictors which minimizes the sum of the squared residuals
- More simply, it is the correlation between the observed values and the fitted values that would be predicted by our model
- Its squared value ( $R^2$ ) is the amount of variance in the dependent variable accounted for by the independent variables

# Multiple Regression: Initial Summary

- So far we have gone through some basics of interpreting a MR analysis
- However an adequate, responsible MR will have much of the analysis spent testing assumptions, identifying problem cases, determining model adequacy, and validating the model
- We will turn to that next

# How does MR work?

- Recall our formula for a straight line in simple regression

$$Y = b_0 + b_1X + e$$

- Now we have multiple predictors, so the formula will now represent vectors and matrices

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

**X** here is the dataframe with a column for the constant

# How does MR work?

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \dots \\ b_k \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{bmatrix}$$

- Dimensions of  $\mathbf{Y} = n, 1$
- Dimensions of  $\mathbf{X} = n, k+1$
- Dimensions for the vector of coefficients  $\mathbf{b} = k+1, 1$
- Dimensions of the vector of residuals  $\mathbf{e} = n, 1$

# How does MR work?

- First we need the coefficients

$$\mathbf{b} = (\mathbf{X}\mathbf{X})^{-1} \mathbf{X}\mathbf{Y}$$

- Take the transpose of the data matrix and multiply it by the original datamatrix, as well as the vector of values of the DV
- Take the inverse of the first product and multiple that by the second product to obtain the vector coefficients
- Multiply the data matrix by the vector of coefficients  $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$
- The result will be a vector of predicted values
- Subtract from the vector of observed to get the vector of residuals

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

# How does MR work?

- Another way:

$$\mathbf{B}_i = \mathbf{R}_{ii}^{-1} \mathbf{R}_{iy}$$

- $\mathbf{B}_i$  = column vector of *standardized* regression coefficients
- $\mathbf{R}_{ii}$  = matrix of correlations among the predictors
- $\mathbf{R}_{iy}$  = column vector of correlations of the DV and predictors

$$\mathbf{R}^2 = \mathbf{R}_{yi} \mathbf{B}_i$$

- $\mathbf{R}_{yi}$  is now the row vector of correlations of the DV and predictors

$$F = \frac{R^2(N - p - 1)}{(1 - R^2)p}$$

$$df = p, N - p - 1$$