Multivariate Models: Model Setup and Assumptions

EPSY 905: Fundamentals of Multivariate Modeling

Online Lecture #10

Today’s Class

- Multivariate linear models: an introduction
- How to form multivariate models in lavaan
- What parameters mean
- How they relate to the multivariate normal distribution

MULTIVARIATE MODELS: AN INTRODUCTION

Multivariate Linear Models

- The next set of lectures are provided to give an overview of multivariate linear models
  - Models for more than one dependent/outcome variable
- Our focus will be on models where the DV is plausibly continuous (so we’ll use error terms that are multivariate normally distributed)
  - Not a necessity – generalized multivariate models are possible

Classical Approaches to Multivariate Linear Models

- In “classical” multivariate textbooks and classes multivariate linear models fall under the names of Multivariate ANOVA (MANOVA) and Multivariate Regression

- These methods rely upon least squares estimation which:
  - Inadequate with missing data
  - Offers very limited methods of setting covariance matrix structures
  - Does not allow for different sets predictor variables for each outcome
  - Does not give much information about model fit
  - Does not provide adequate model comparison procedures

- The classical methods have been ***subsumed*** into the modern (likelihood or Bayes-based) multivariate methods
  - ***Subsume:*** include or absorb (something) in something else
  - Meaning: modern methods do what classical methods do (and more)

Contemporary Methods for Estimating Multivariate Linear Models

- We will discuss three large classes of multivariate linear modeling methods:
  - Path analysis models (typically through structural equation modeling and path analysis software)
  - Linear mixed models (typically through linear models software)
  - Bayesian networks (frequently not mentioned in social sciences but subsume all we are doing)
- The theory behind each is identical – the main difference is in software
  - Some software does a lot (Mplus is likely the most complete), but none (as of March 2018) do it all
- We will start with path analysis (via the lavaan package) as the modeling method is more direct but then move to linear mixed models software (via the nlme and lme4 packages) to be complete in our discussion
- Bayesian networks will be discussed in the Bayes section of the course and will use entirely different software

Planned Course Outline

- We will start with path analysis (via the lavaan package) as the modeling method is more direct

- but then move to linear mixed models software (via the nlme and lme4 packages) to be complete in our discussion

- Bayesian networks will be discussed in the Bayes section of the course and will use entirely different software

- The frustrating part of each method is that each relies upon different estimation methods
  - So results sometimes lack comparability ‽‽‽

The Curse of Dimensionality: Shared Across Models

- Having lots of parameters creates a number of problems
  - Estimation issues for small sample sizes
  - Power to detect effects
  - Model fit issues for large numbers of outcomes

- For multivariate normal data: having a quadratic increase in the number of parameters as the number of outcomes increases linearly is sometimes called the “curse of dimensionality”

- To be used as an analysis model, however, a covariance structure must “fit” **as well as** the **saturated/unstructured** covariance matrix

Biggest Difference From Univariate Models: Model Fit

- In univariate linear models the “model for the variance” wasn’t much of a model
  - There was one variance term possible and one term estimated
    - A saturated model
  - Model fit was always perfect

- Because of the number of variances/covariances, multivariate models often don’t have saturated models for the variances
  - Therefore, model fit becomes an issue

- Any non-saturated model for the variances must be shown to fit the data** before being used for interpretation
  - ** fit the data has differing standards depending on software type used

EXAMPLE DATA SET

Today's Data Example

- Data are simulated based on the results reported in:

Pajares, F., & Miller, M. D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: a path analysis. *Journal of Educational Psychology, 86*, 193-203.

- Sample of 350 undergraduates (229 women, 121 men)
  - In simulation, 10% of variables were missing (using missing completely at random mechanism)

- Note: simulated data characteristics differ from actual data (some variables extend beyond their official range)
  - Simulated using Multivariate Normal Distribution
    - Some variables had boundaries that simulated data exceeded
  - Results will not match exactly due to missing data and boundaries

Variables of Data Example

- Female (sex variable: 0 = male; 1 = female)
- Math Self-Efficacy (MSE)
  - Reported reliability of .91
  - Assesses math confidence of college students
- Perceived Usefulness of Mathematics (USE)
  - Reported reliability of .93
- Math Anxiety (MAS)
  - Reported reliability ranging from .86 to .90
- Math Self-Concept (MSC)
  - Reported reliability of .93 to .95
- Prior Experience at High School Level (HSL)
  - Self report of number of years of high school during which students took mathematics courses
- Prior Experience at College Level (CC)
  - Self report of courses taken at college level
- Math Performance (PERF)
  - Reported reliability of .788
  - 18-item multiple choice instrument (total of correct responses)

MAXIMUM LIKELIHOOD BY MVN: USING LAVAAN FOR ESTIMATION

Using MVN Likelihoods in lavaan

- Lavaan's default model is a linear (mixed) model that uses ML with the multivariate normal distribution

- ML is sometimes called a full information method (FIML)
  - Full information is the term used when each observation gets used in a likelihood function
  - The contrast is limited information (not all observations used; typically summary statistics are used)

- You can use lavaan to do analyses for all sorts of linear models including:
  - MANOVA
  - Repeated Measures ANOVA
  - Factor Models

Revisiting Univariate Linear Regression

- We will begin our discussion by starting with perhaps the simplest model we will see: a univariate empty model
  - We will use the PERF variable from our example data: Performance on a mathematics assessment

- The empty model for PERF is:

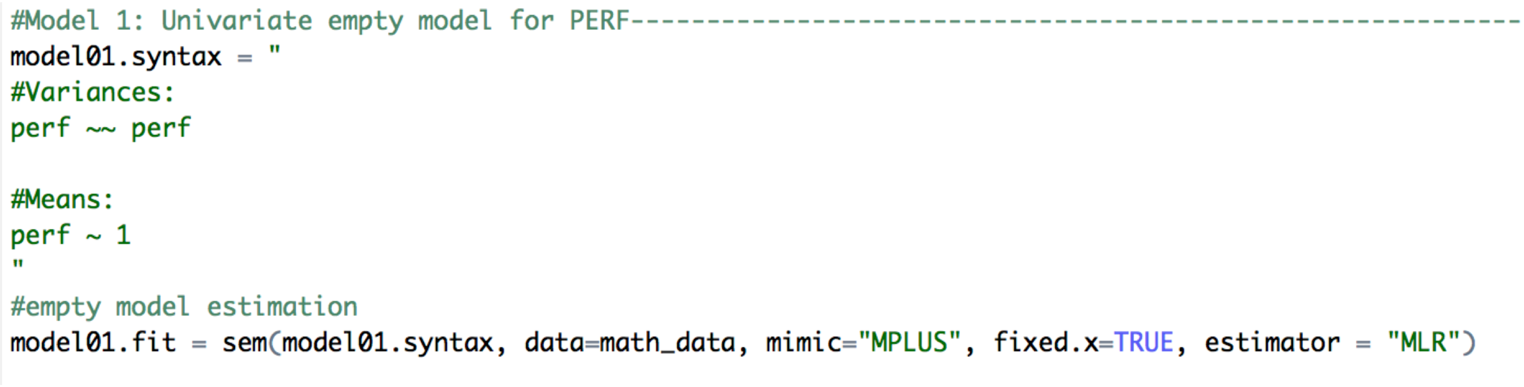
$$PERF_i = \beta_{0,PERF} + e_{i,PERF}$$

  - Additionally,$e_{i,PERF} \sim N\left(0, \sigma^2_{e,PERF}\right)$
  - So, two parameters are estimated: $\beta_{0,PERF}$ and $\sigma^2_{e,PERF}$

- Here, the additional subscript is added to denote these terms are part of the model for the PERF variable
  - We will need these when we get to multivariate models and path analysis

lavaan Syntax

- The lavaan package works by taking the typical R model syntax (as from lm()) and putting it into a quoted character variable
  - lavaan model syntax also includes other commands used to access other parts of the model (really, parts of the MVN distribution)
- Here: ~~ indicate variance or covariance between variables on either side (perf ~~ perf) estimates variance
  - perf ~ 1 estimates intercept for perf

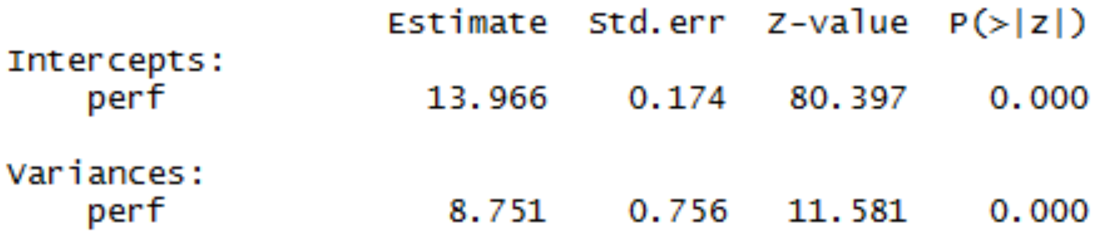


```
#Model 1: Univariate empty model for PERF-----------------------------------------------------------
model01.syntax = "
#Variances:
perf ~~ perf

#Means:
perf ~ 1
"
#empty model estimation
model01.fit = sem(model01.syntax, data=math_data, mimic="MPLUS", fixed.x=TRUE, estimator = "MLR")
```

- We use the sem() function to run the model

Model Parameter Estimates and Assumptions

- Interpret the following estimates…$\beta_{0,PERF}$ and $\sigma^2_{e,PERF}$



```
                    Estimate  Std.err  Z-value  P(>|z|)
Intercepts:
    perf              13.966    0.174   80.397    0.000

Variances:
    perf               8.751    0.756   11.581    0.000
```

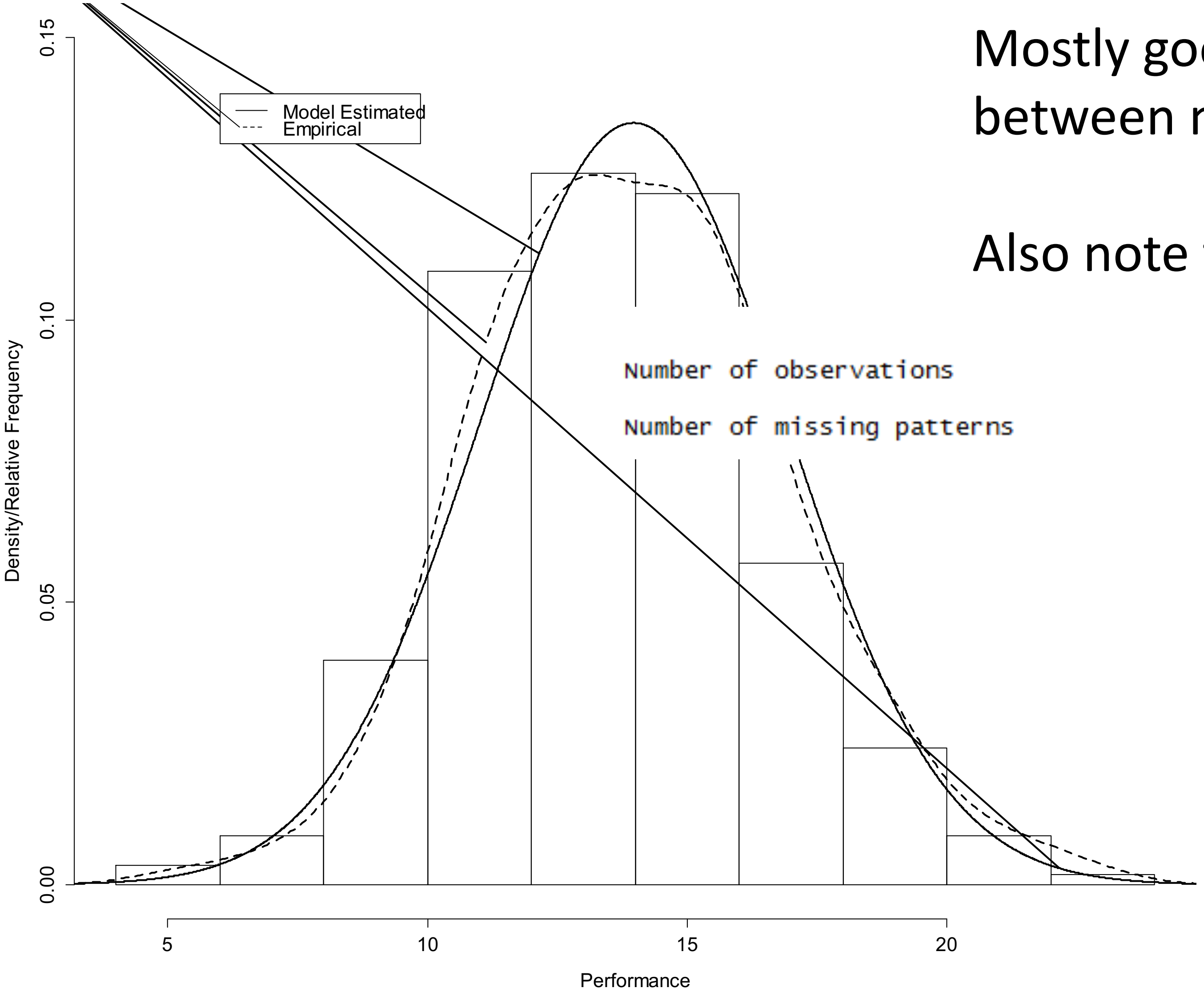
- As we assumed $e_{i,PERF} \sim N\left(0, \sigma^2_{e,PERF}\right)$ we assume the following about the data:

$$PERF_i \sim N\left(\beta_{0,PERF}, \sigma^2_{e,PERF}\right)$$

- Using the model estimates, this becomes:

$$PERF_i \sim N(13.966, 8.751)$$

Plot of Model Estimated vs. Data



Mostly good agreement between model and data

Also note the sample size:

```
                                                      Used          Total
Number of observations                                 290            350

Number of missing patterns                               1
```

MULTIVARIATE EMPTY MODELS

Adding One More Variable: Multivariate Regression

- We will now move to modeling two variables from our example data that we wish to describe:
  - Mathematics performance (PERF)
  - Perceived usefulness (PERF)
- We will assume these to be continuous variables (conditionally MVN)
- Initially, we will only look at an empty model with these two variables
  - Empty models are baseline models
  - We will use these to show how such models look based on the characteristics of the multivariate normal distribution
  - We will also show the bigger picture when modeling multivariate data: how we must be sure to model the covariance matrix correctly

Multivariate Empty Model: The Notation

- The multivariate model for PERF and USE is given by two regression models, which are estimated simultaneously:

$$PERF_i = \beta_{0,PERF} + e_{i,PERF}$$
$$USE_i = \beta_{0,USE} + e_{i,USE}$$

- As there are two variables, the error terms have a joint distribution that will be multivariate normal:

$$\begin{bmatrix} e_{i,PERF} \\ e_{i,USE} \end{bmatrix} \sim N_2\left(\mathbf{0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \sigma^2_{e,PERF} & \sigma_{e,PERF,USE} \\ \sigma_{e,PERF,USE} & \sigma^2_{e,USE} \end{bmatrix}\right)$$

- Each error term has its own variance but now there is a covariance between error terms
  - We will soon see that the overall **R** matrix structure can be modified

Data Model

- Before showing the syntax and the results, we must first describe how the multivariate empty model implies how our data should look

- Multivariate model with matrices:

$$\begin{bmatrix} PERF_i \\ USE_i \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{0,PERF} \\ \beta_{0,USE} \end{bmatrix} + \begin{bmatrix} e_{i,PERF} \\ e_{i,USE} \end{bmatrix}$$

$$\mathbf{Y}_i = \mathbf{X}_i \mathbf{B} + \mathbf{e}_i$$

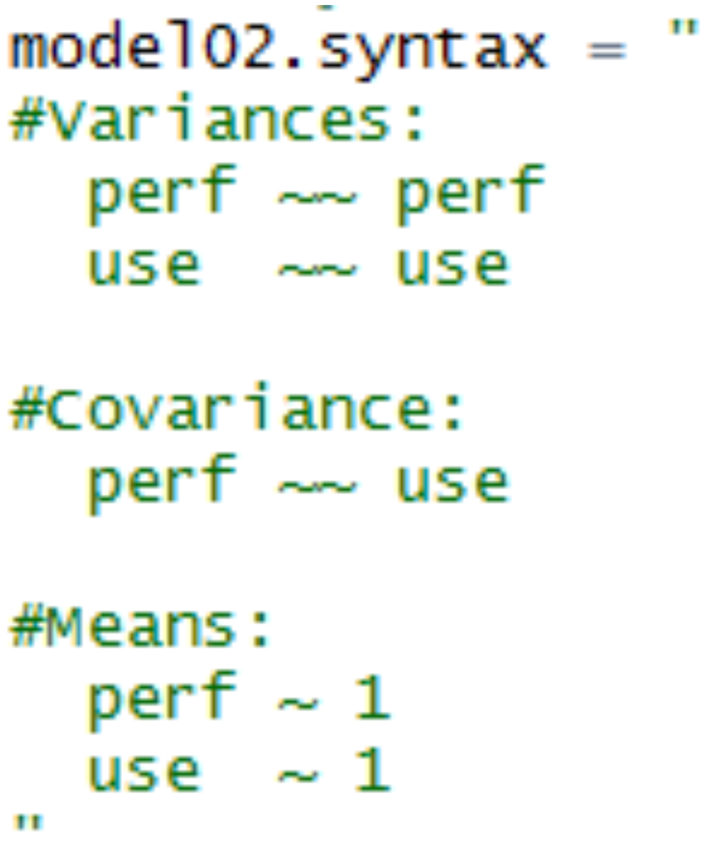
- Using expected values and linear combination rules, we can show that:

$$\mathbf{Y}_i \sim N_2(\mathbf{X}_i \mathbf{B}, \mathbf{V}_i)$$

$$\begin{bmatrix} PERF_i \\ USE_i \end{bmatrix} \sim N_2\left(\boldsymbol{\mu}_i = \begin{bmatrix} \beta_{0,PERF} \\ \beta_{0,USE} \end{bmatrix}, \mathbf{V}_i = \mathbf{R} = \begin{bmatrix} \sigma^2_{e,PERF} & \sigma_{e,PERF,USE} \\ \sigma_{e,PERF,USE} & \sigma^2_{e,USE} \end{bmatrix} \right)$$

Lavaan Multivariate Regression Model Syntax

$$\begin{bmatrix} PERF_i \\ USE_i \end{bmatrix} \sim N_2\left(\begin{bmatrix} \beta_{0,PERF} \\ \beta_{0,USE} \end{bmatrix}, \begin{bmatrix} \sigma^2_{e,PERF} & \sigma_{e,PERF,USE} \\ \sigma_{e,PERF,USE} & \sigma^2_{e,USE} \end{bmatrix}\right)$$



```
model02.syntax = "
#Variances:
  perf ~~ perf
  use  ~~ use

#Covariance:
  perf ~~ use

#Means:
  perf ~ 1
  use  ~ 1
"
```

$\sigma^2_{e,PERF}$

$\sigma^2_{e,USE}$

$\sigma_{e,PERF,USE}$

$\beta_{0,PERF}$

$\beta_{0,USE}$

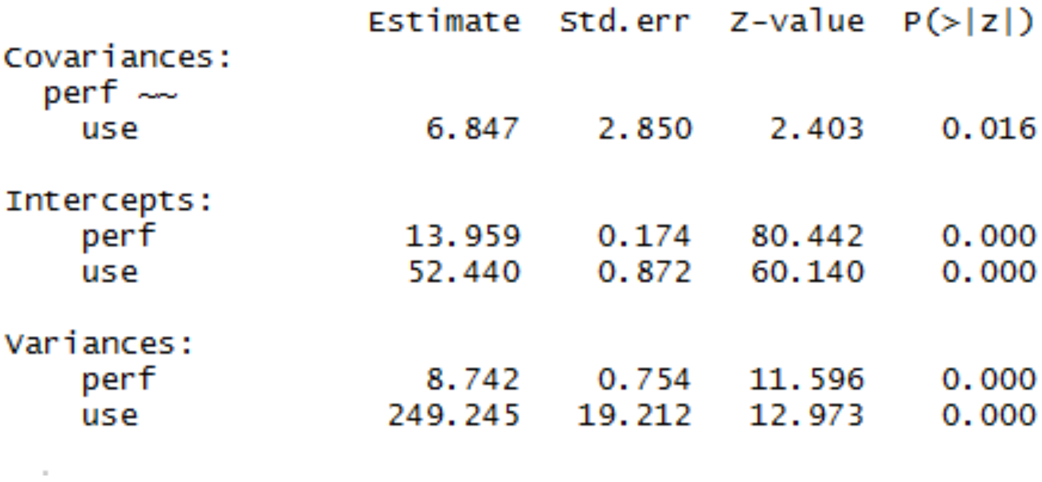
This covariance matrix is said to be **saturated**: All parameters are estimated

It is also called an **unstructured** covariance matrix

No other structure for the covariance matrix can fit better (only as well as)

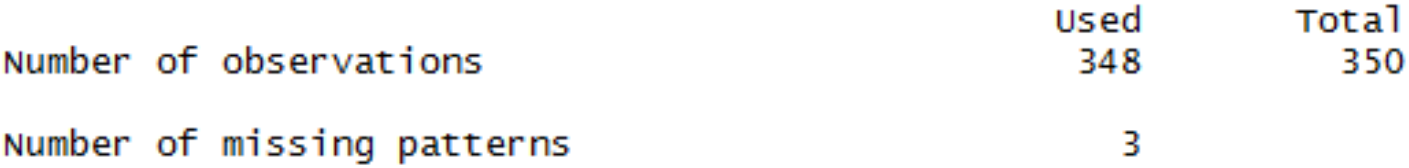
Multivariate Regression Model Results

- The estimated values:
  - What is the estimated correlation between PERF and USE?



```
                      Estimate  Std.err  Z-value  P(>|z|)
Covariances:
  perf ~~
    use                  6.847    2.850    2.403    0.016

Intercepts:
    perf                13.959    0.174   80.442    0.000
    use                 52.440    0.872   60.140    0.000

Variances:
    perf                 8.742    0.754   11.596    0.000
    use                249.245   19.212   12.973    0.000
```

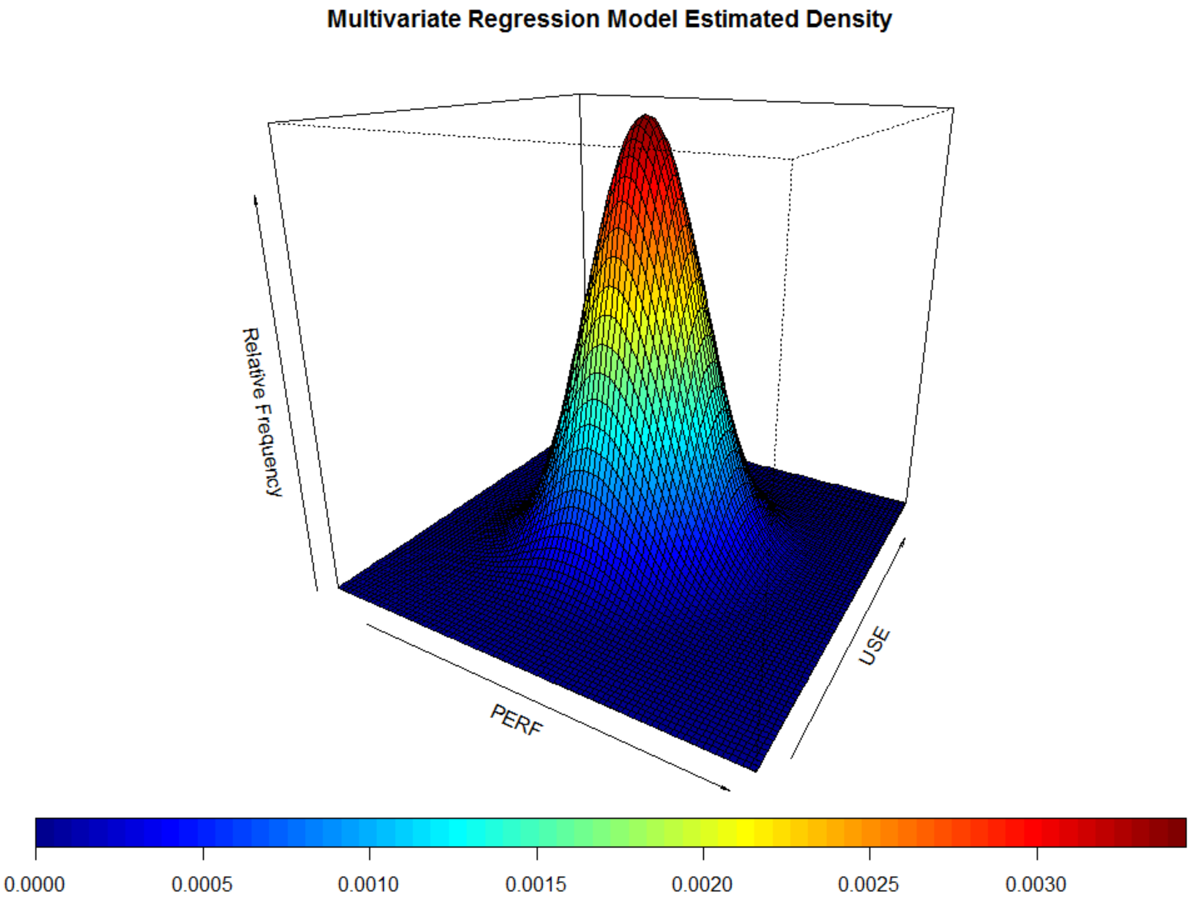
- [Side Note] Why is the sample size different from Model 1?



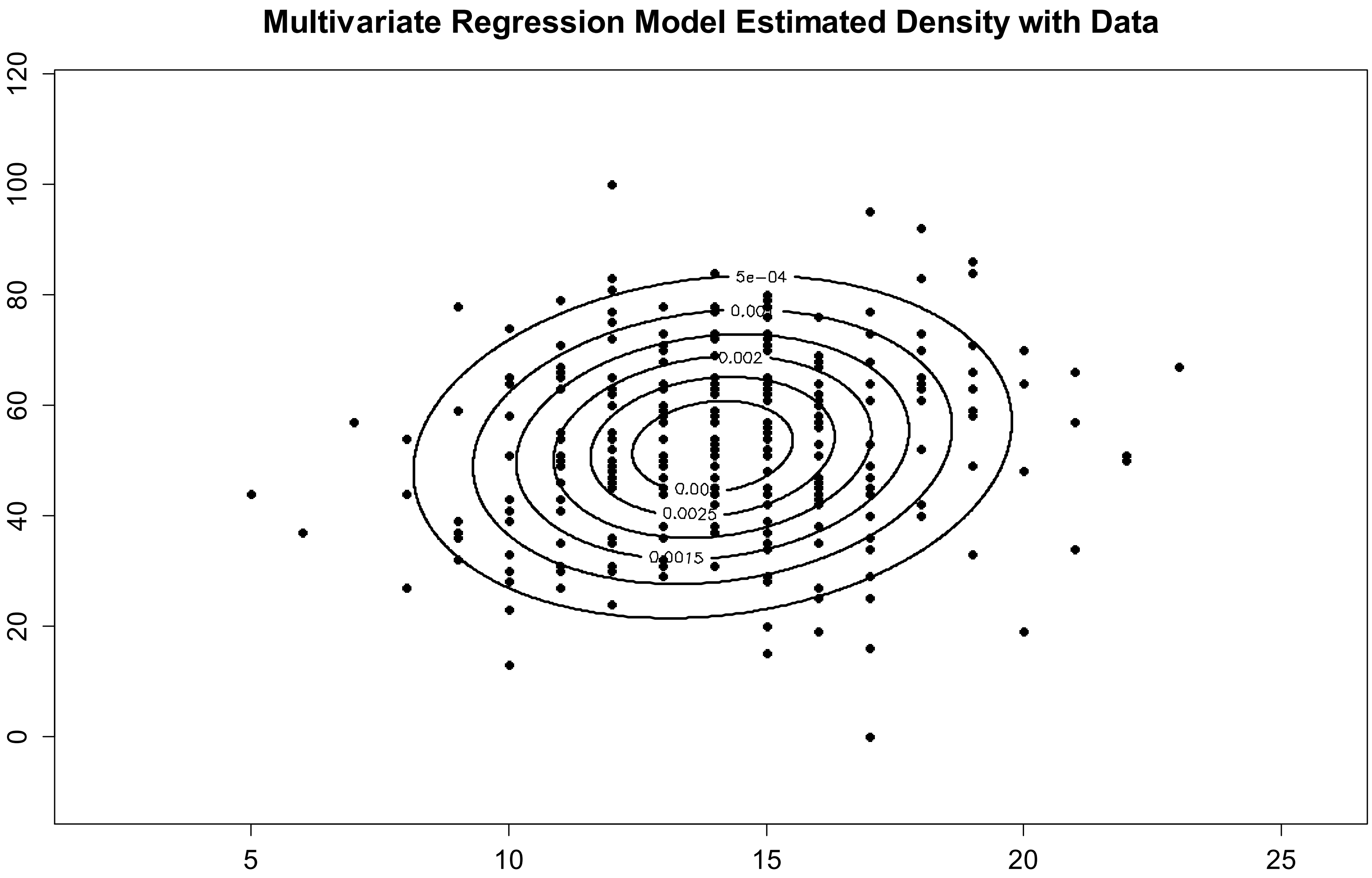
```
                                              Used       Total
Number of observations                         348         350

Number of missing patterns                       3
```

Plotting the Model Estimated Results

$$\begin{bmatrix} PERF_i \\ USE_i \end{bmatrix} \sim N_2\left(\begin{bmatrix} 13.959 \\ 52.440 \end{bmatrix}, \begin{bmatrix} 8.742 & 6.847 \\ 6.847 & 249.245 \end{bmatrix}\right)$$

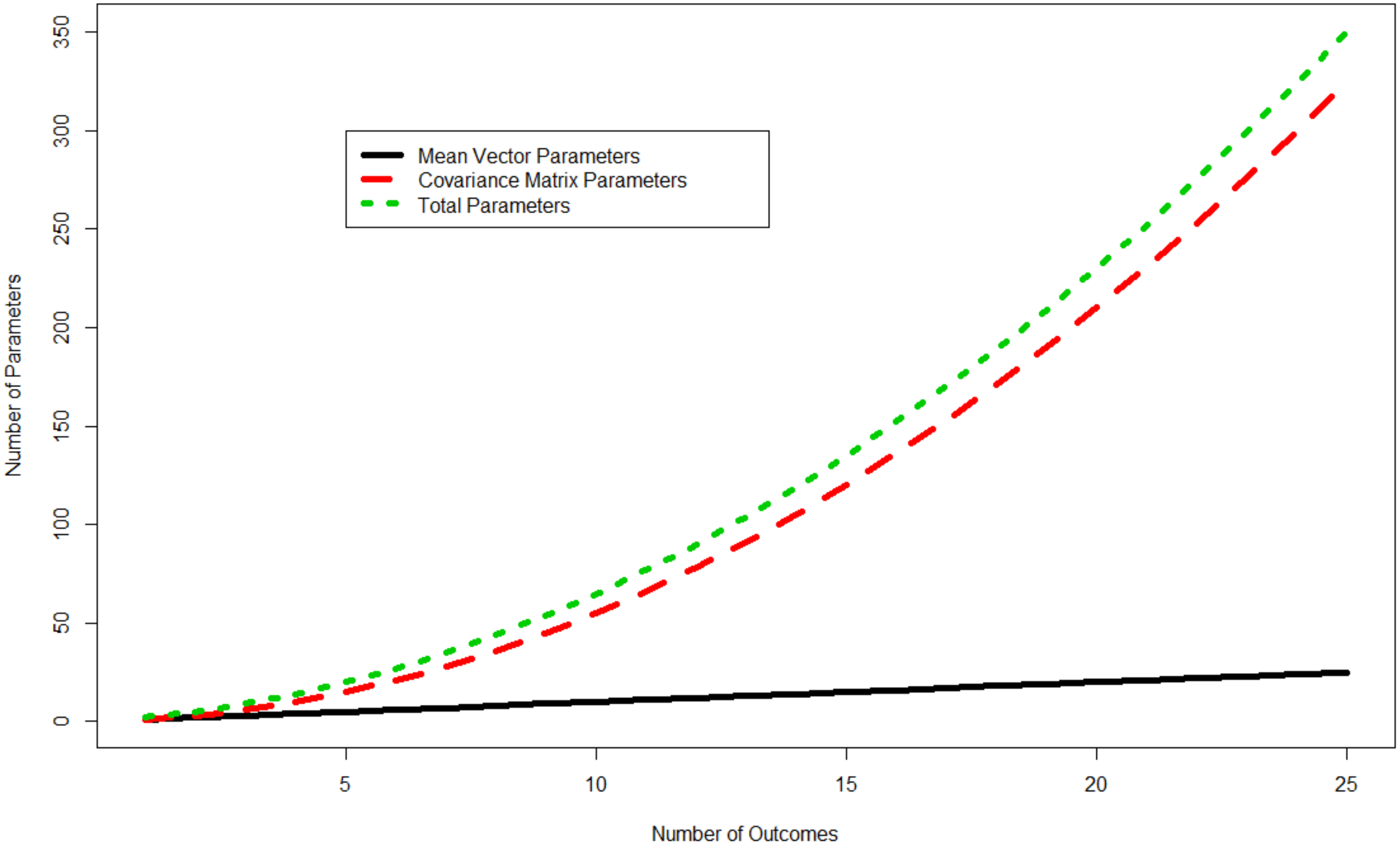


Comparing Model with Data



The Problem with Multivariate Models

- As more dependent variables (outcomes) are added to a multivariate model, the number of parameters needed for a saturated model gets very large:



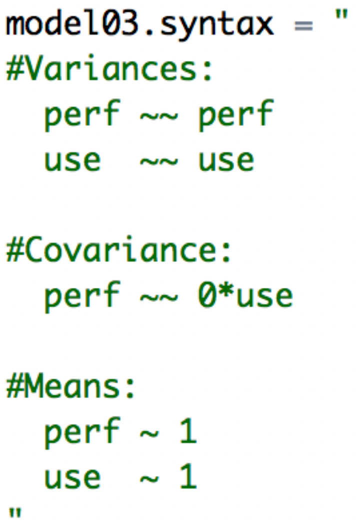
A SECOND MODEL

A Different Model for the Data

- To demonstrate how models may vary in terms of model fit (and to set up a discussion of model fit and model comparisons) we will estimate a model where we set the covariance between PERF and USE to zero
  - Zero covariance implies zero correlation – which is unlikely to be true given our previous analysis
- You likely would not use this model in a real data analysis
  - If anything, you may start with a zero covariance and then estimate one
- But, this will help to introduce come concepts needed to assess the quality of the multivariate model

Lavaan Syntax

- The lavaan() syntax for setting the covariance to zero is:



```
model03.syntax = "
#Variances:
  perf ~~ perf
  use  ~~ use

#Covariance:
  perf ~~ 0*use

#Means:
  perf ~ 1
  use  ~ 1
"
```

- The only difference is 0* in front of USE in the ~~ section
  - You can set a parameter equal to a value this way

Model Assumptions

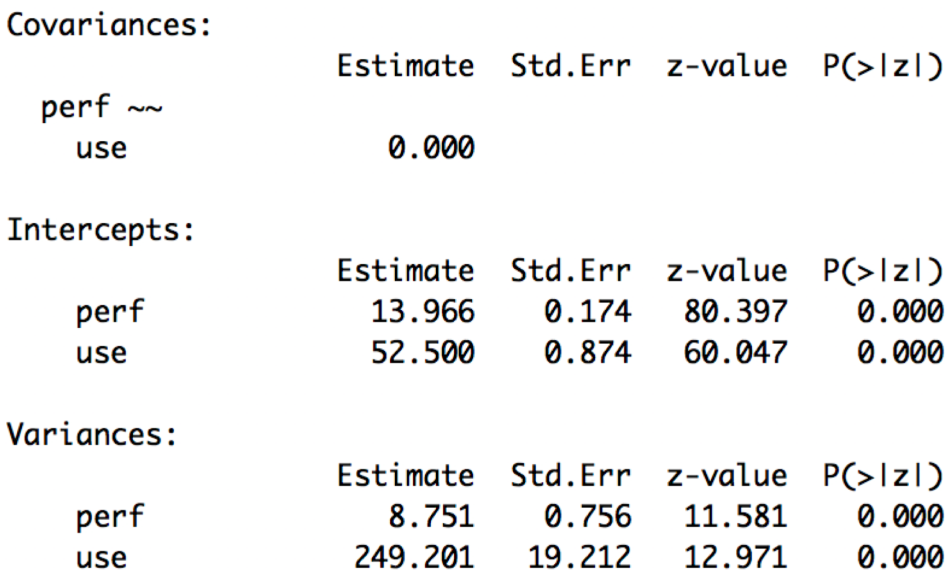
- The zero covariance now leads to the following assumptions about the data:

$$\begin{bmatrix} PERF_i \\ USE_i \end{bmatrix} \sim N_2\left(\begin{bmatrix} \beta_{0,PERF} \\ \beta_{0,USE} \end{bmatrix}, \begin{bmatrix} \sigma^2_{e,PERF} & 0 \\ 0 & \sigma^2_{e,USE} \end{bmatrix}\right)$$

- Because these are MVN, we are assuming PERF is *independent* from USE (has zero correlation/covariance)

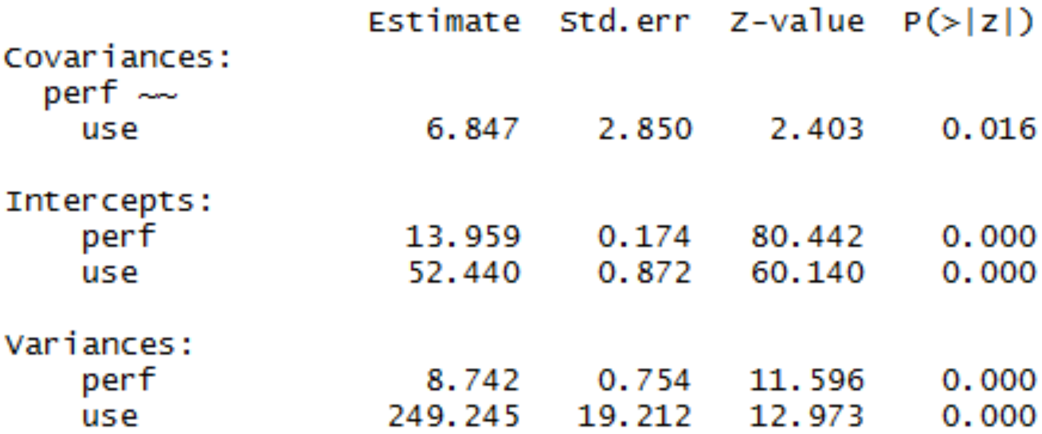
Model Results

Model 3



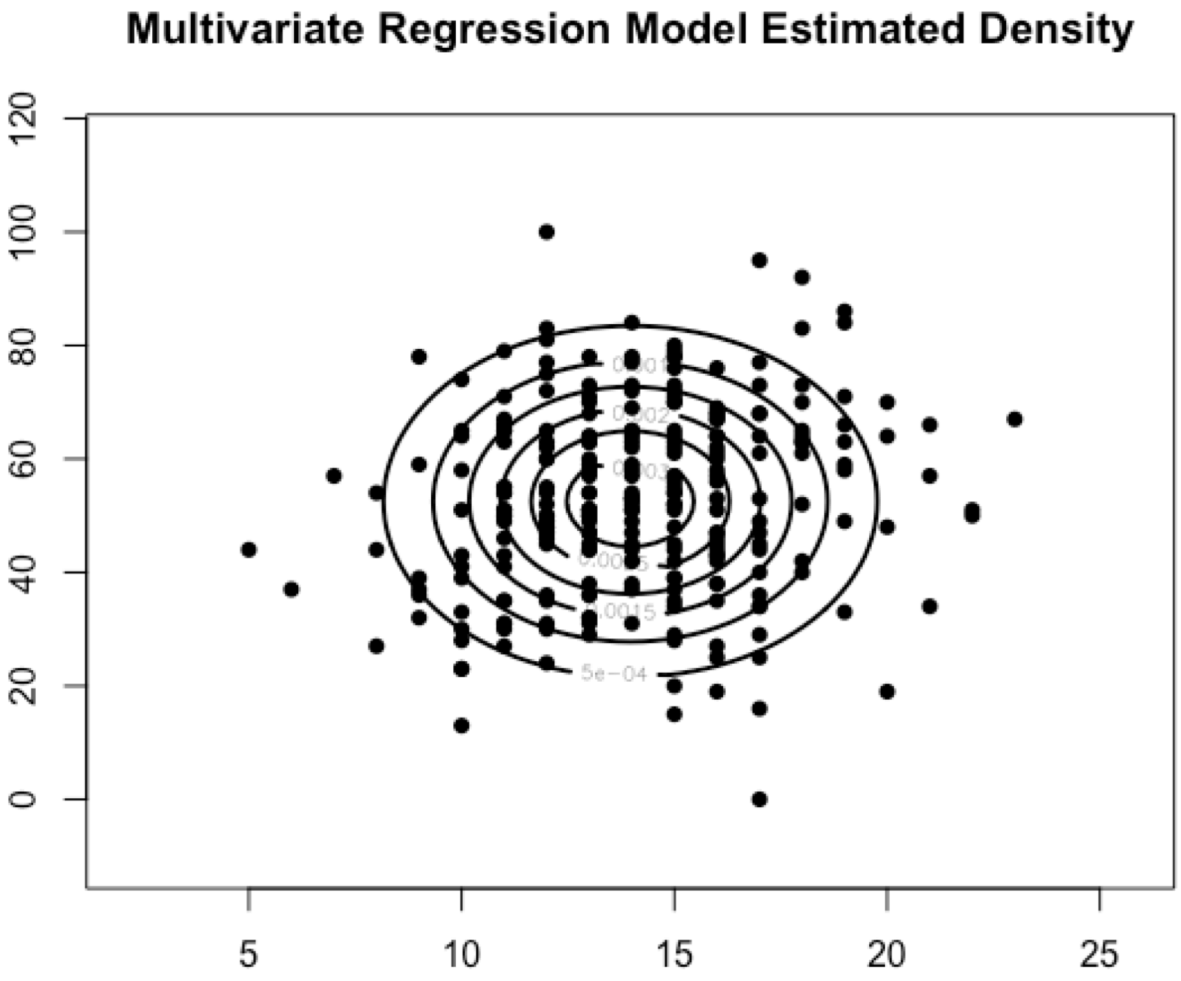
```
Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
  perf ~~
    use               0.000

Intercepts:
                   Estimate  Std.Err  z-value  P(>|z|)
    perf             13.966    0.174   80.397    0.000
    use              52.500    0.874   60.047    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    perf              8.751    0.756   11.581    0.000
    use             249.201   19.212   12.971    0.000
```

Model 2



```
                   Estimate  Std.err  Z-value  P(>|z|)
Covariances:
  perf ~~
    use               6.847    2.850    2.403    0.016

Intercepts:
    perf             13.959    0.174   80.442    0.000
    use              52.440    0.872   60.140    0.000

Variances:
    perf              8.742    0.754   11.596    0.000
    use             249.245   19.212   12.973    0.000
```

Examining Model/Data Fit



Questions Remain

- Does each model fit the data well (absolute model fit)?
  - If not, how can we improve model fit?

- Which model fits better (relative model fit)?

- Answers are given in the following lectures…

WRAPPING UP

Wrapping Up

- This lecture was an introduction to the estimation of multivariate linear models for multivariate outcomes the using path analysis/SEM package lavaan

- We saw that the model for continuous data uses the multivariate normal distribution in its likelihood function