# Lecture 2: Latent Variable Models

Jeff Rouder

June, 2024

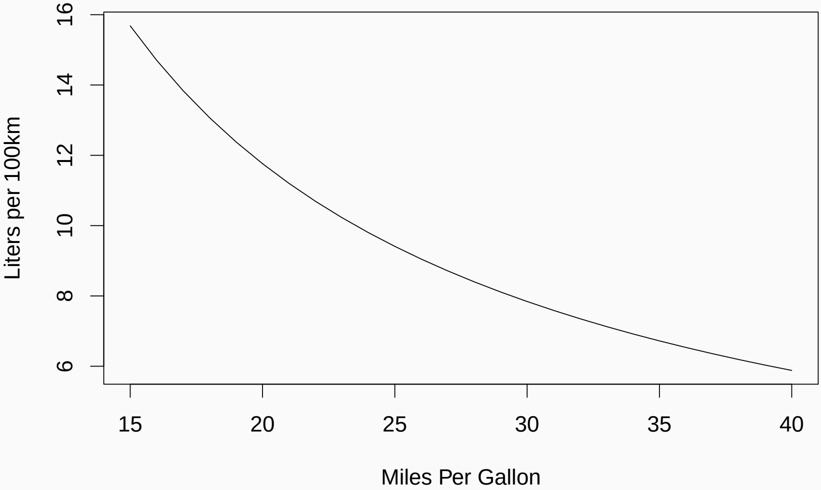
# Meet Carl

# Fuel Efficiency

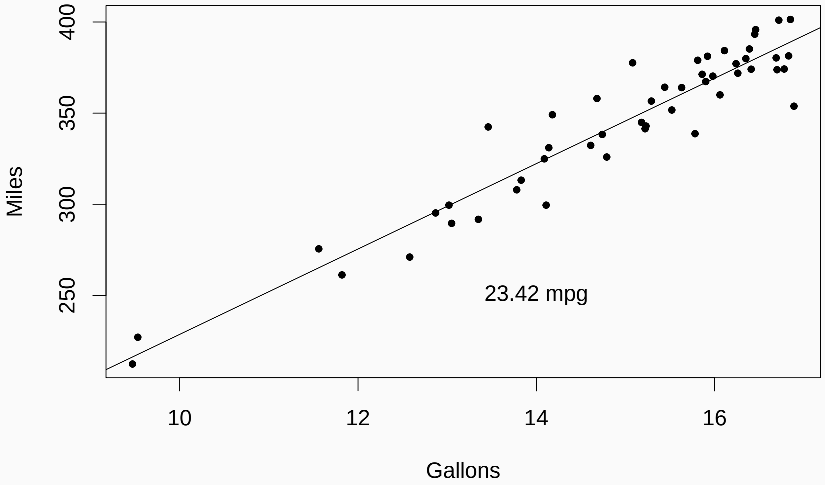
- In U.S., how many miles per gallon? $m$
- In Europe, how many liters per 100km? $\ell$

$$\ell = \frac{1}{m} \times \frac{1}{1.609} \times \frac{3.785}{1} \times 100$$
$$\ell = \frac{235.24}{m}$$

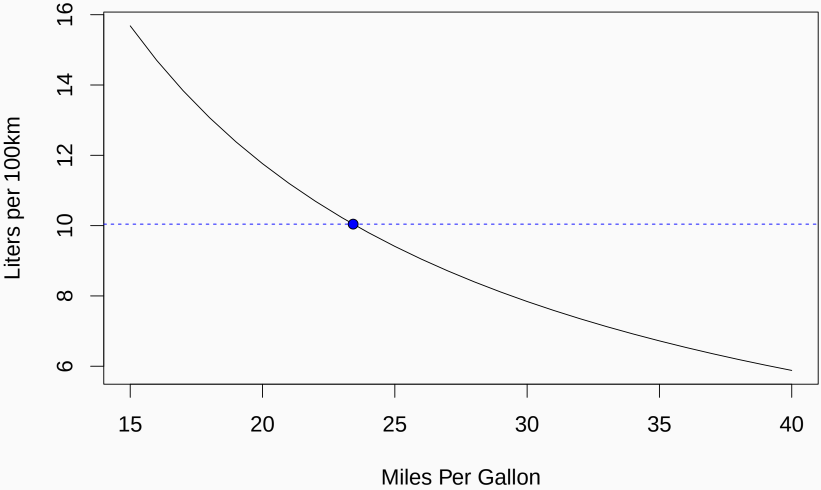
# Fuel Efficiency



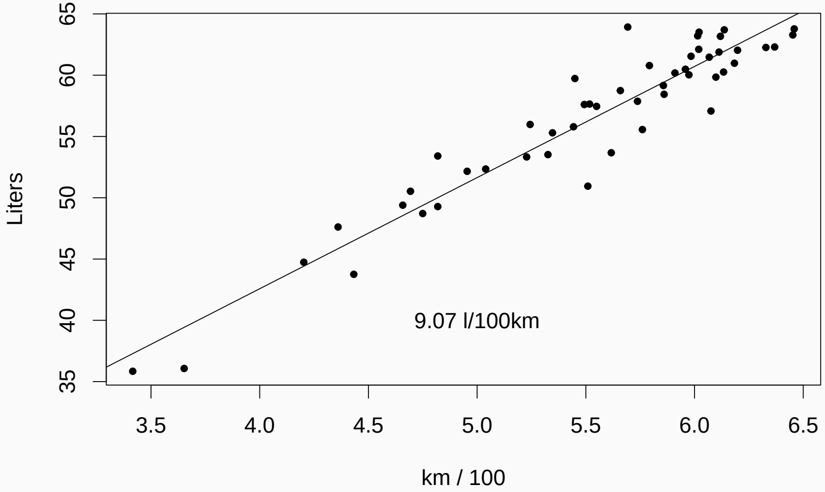
# My Car Over 50 Fills



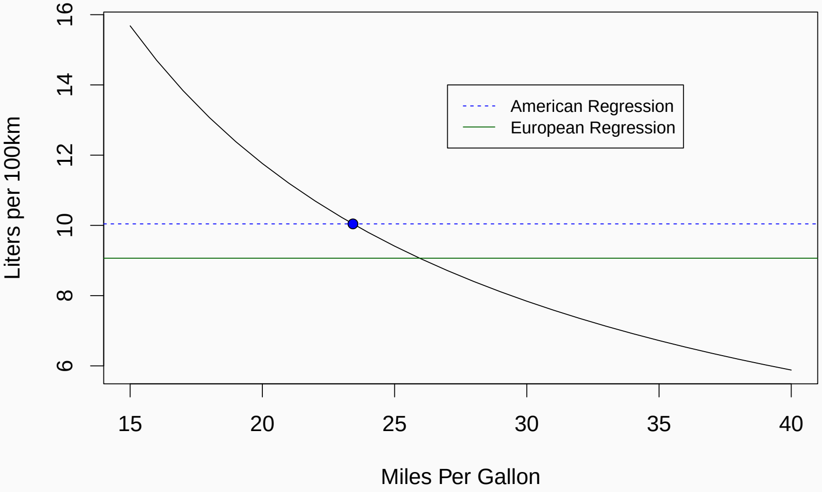
# Fuel Efficiency of My Car



# My Car, European Approach



# Paradox



American Regression
European Regression
Liters per 100km
Miles Per Gallon
16
14
12
10
8
6
15
20
25
30
35
40

# Upshot

- The data are fixed
- 10% difference in fuel efficiency is unexpected and troubling
- The 10% difference is systematic. Happens in the limit of infinite tank fills.
- Why?
  - Regressing X on Y is not the same as regressing Y on X.

# Resolution

- OLS is conditional. We are understanding Y given X. $Y|X$ is not the same as $X|Y$.
- To get the structural relations among X and Y, we might wish to study the joint distribution without conditioning on one or the other.
- Modeling random variables jointly—multivariate modeling—is often aided with the introduction of latent variables to account for correlation.

# Bivariate Normal

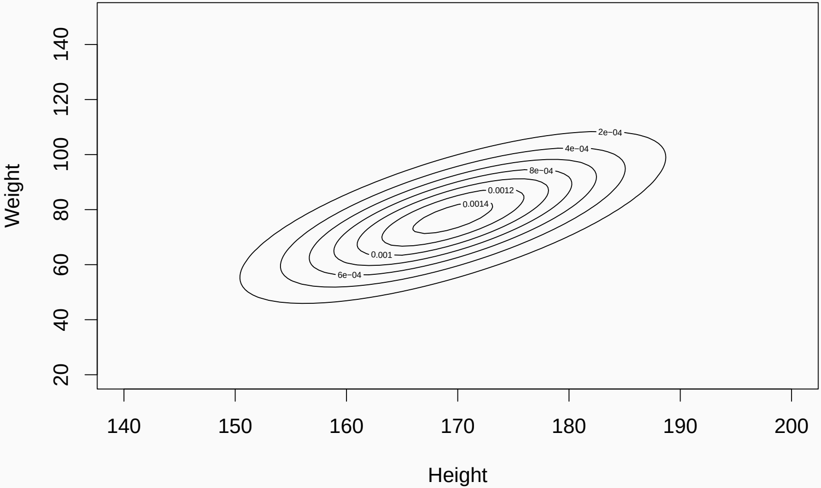
- $h$: Height
- $w$: Weight

$$\begin{pmatrix} h \\ w \end{pmatrix} \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

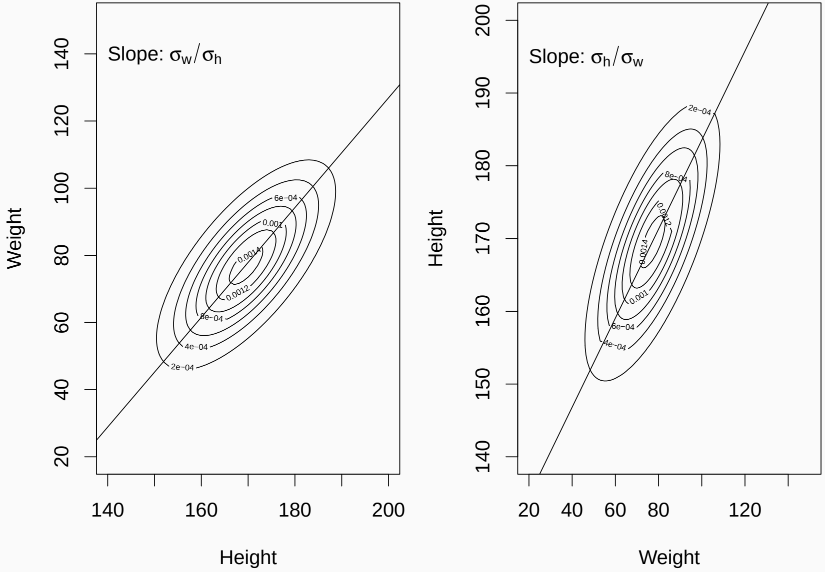
or

$$\begin{pmatrix} h \\ w \end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix} \mu_h \\ \mu_w \end{pmatrix}, \begin{pmatrix} \sigma_h^2 & \rho\sigma_h\sigma_w \\ \rho\sigma_h\sigma_w & \sigma_w^2 \end{pmatrix}\right)$$

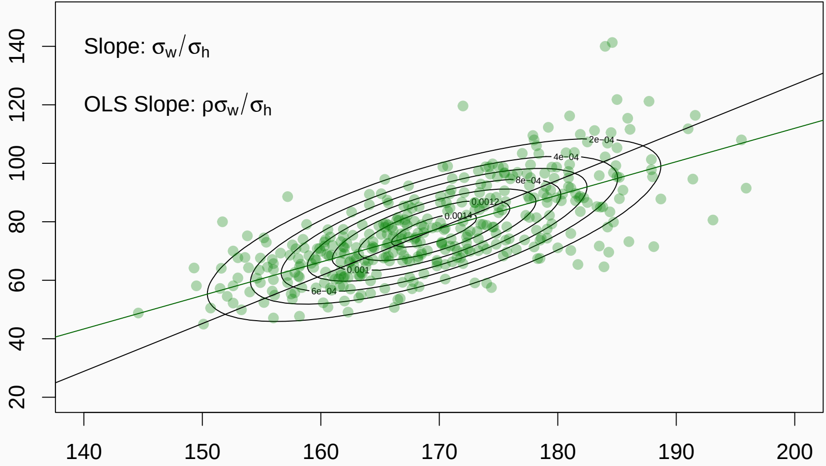
# Bivariate Normal



# Bivariate Normal



# Real Height and Weight



# Notation for many observations

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{1J} \end{pmatrix} \sim \mathrm{N}_J \left( \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1J}\sigma_1\sigma_J \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \dots & \rho_{2J}\sigma_2\sigma_J \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1J}\sigma_1\sigma_J & \rho_{2J}\sigma_2\sigma_J & \dots & \sigma_J^2 \end{pmatrix} \right)$$

or

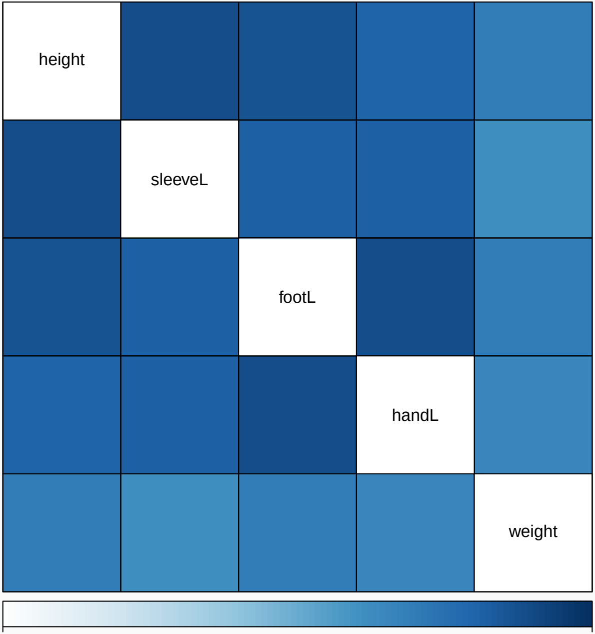
$$\boldsymbol{Y}_i \sim \mathrm{N}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- All latent variable models are **constraints** on $\boldsymbol{\Sigma}$.

# Real Data From 5 Measures of Body Size

```
##         height sleeveL footL handL weight
## height   1.000   0.888 0.868 0.810  0.700
## sleeveL  0.888   1.000 0.818 0.810  0.617
## footL    0.868   0.818 1.000 0.887  0.693
## handL    0.810   0.810 0.887 1.000  0.650
## weight   0.700   0.617 0.693 0.650  1.000
```

# Real Data 5 Measures



# General Model

- How many unique correlations are there in 5 tasks?
- $J(J-1)/2 = 10$.
- No more than 10 correlations, 5 variances, 5 means.

# One Factor Model

- Let $i = 1, \ldots, I$ denote people
- Let $j = 1, \ldots, J$ denote scores (items, measures, tasks)

$$Y_{ij} | \phi_i \sim \mathrm{N}(\mu_j + \lambda_j \phi_i, \delta_j^2)$$
$$\phi_i \sim \mathrm{N}(0, 1)$$

- $\phi_i$ is **factor score** for the *i*th person
    - latent ability
- $\lambda_j$ is **factor loading** for *j*th score

# Marginal Model

- In frequentist analysis, $\phi_i$ has a nuanced (perhaps nonsensical) status. It is neither a datum (not observed) nor a parameter (has a distribution). It is *latent datum*.
- Often, it is marginalized across. **Marginalization** is really, really helpful.

# Marginalization

- $f(y) = \int_\theta f(y \mid \theta) d\theta$
- There are shortcuts for the normal

# Marginal Model

Conditional Model

$$\begin{aligned} Y_{ij}|\phi_i &\sim \mathrm{N}(\mu_j + \lambda_j \phi_i, \delta_j^2) \\ \phi_i &\sim \mathrm{N}(0, 1) \end{aligned}$$

Marginal Model

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{iJ} \end{pmatrix} \sim \mathrm{N}_J \left( \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix}, \begin{pmatrix} \delta_1^2 + \lambda_1^2 & \lambda_1\lambda_2 & \dots & \lambda_1\lambda_J \\ \lambda_1\lambda_2 & \delta_2^2 + \lambda_2^2 & \dots & \lambda_2\lambda_J \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1\lambda_J & \lambda_2\lambda_J & \dots & \delta_J^2 + \lambda_J^2 \end{pmatrix} \right)$$

# Full Matrix Notation

Conditional Model

$$\boldsymbol{Y}_i|\phi_i \sim \mathrm{N}(\boldsymbol{\mu} + \boldsymbol{\lambda}\phi_i, D(\boldsymbol{\delta}^2))$$
$$\phi_i \sim \mathrm{N}(0, 1)$$

- Note: $D(\boldsymbol{\delta}^2)$ mean diagonal matrix with diagonal $\boldsymbol{\delta}^2$.

Marignal Model

$$\boldsymbol{Y}_i \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\lambda}\boldsymbol{\lambda}' + D(\boldsymbol{\delta}^2))$$

# Running in Lavaan

- Why I like lavaan
    - great documentation
    - fast
    - lots of options

## Running in Lavaan

```r
#install.packages('lavaan')
library(lavaan)
fit=efa(data=y,nfactors=1,rotation="varimax")
summary(fit)
```

```
## This is lavaan 0.6.17 -- running exploratory factor anal
##
##   Estimator                                          ML
##   Rotation method                    VARIMAX ORTHOGONAL
##   Rotation algorithm (rstarts)                 GPA (30)
##   Standardized metric                              TRUE
##   Row weights                                    Kaiser
##
##   Number of observations                            400
##
## Fit measures:
```

# What Does This Output Mean?

- Factor loadings are standardized. What does that mean?
- Two main types of standardization
  - Effect-size
  - Total

## Standardization

$$Y_i = \mu + X_i\theta + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- Assume centered covariates. Mean of $X_i = 0$
- What are the units?
    - $Y_i$: $u_y$
    - $\mu$: $u_y$
    - $X_i$: $u_x$
    - $\sigma^2$: $u_y^2$
    - $\theta$: $u_y/u_x$. Also $-\infty < \theta < \infty$

# Standardization of covariates

- Standardize covariate:
  - $m_x = \sum(X_i)/N, \quad s_x = \sqrt{\sum X_i^2/N}$
  - $W_i = (X_i - m_x)/s_x$

$$Y_i = \mu + W_i\theta + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- $W_i$ is unitless
- $\theta$ is in units $u_y$

# Standardization, Residual

$$Y_i = \mu + \sigma W_i \theta^\dagger + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- $\theta^\dagger$ is effect size, standardized w/ respect to variability in both covariate and residual.
- $\theta^\dagger = \theta/\sigma$
- $\theta^\dagger$ is unitless, still $-\infty < \theta^\dagger < \infty$

$$\frac{Y_i - \mu}{\sigma} = W_i \theta^\dagger + \epsilon_i^\dagger, \quad \epsilon_i^\dagger \sim \mathrm{N}(0, 1)$$

# Standardization, Full Variance

$$Y_i = \mu + \sigma_y \theta^* + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2)$$

- $\theta^*$ is an effect size too, but it is standardized w/ respect to variability in covariate and the dependent measure.
- $\theta^* = \theta / \sigma^y$
- $\sigma_y^2 = \theta^2 + \sigma^2$
- $(\theta^*)^2 = \theta^2 / (\sigma^2 + \theta^2) = \rho^2$
- $-1 \leq \theta^* \leq 1$

$$\frac{Y_i - \mu}{\sigma_y} = W_i \theta^* + \epsilon_i^*, \quad \epsilon_i^* \sim \mathrm{N}\left( (0, \frac{\sigma^2}{\sigma^2 + \theta^2} \right)$$

# Factor Model Version

$$\lambda_j^* = \frac{\lambda_j}{\sqrt{\lambda_j^2 + \delta_j^2}}$$

- Standardized Loadings: $0 \leq \lambda_j \leq 1$

# For Any Covariance:

$$
\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_J \end{pmatrix} \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1J} \\ \rho_{12} & 1 & \dots & \rho_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1J} & \rho_{2J} & \dots & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_J \end{pmatrix}
$$

$$
= D(\boldsymbol{\sigma})\boldsymbol{\rho}D(\boldsymbol{\sigma})
$$

- show it with a two-by-two example.

# One-Factor Model

- $\boldsymbol{\Sigma} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + D(\boldsymbol{\delta}^2)$
- $D(\boldsymbol{\Sigma}^2) = D(\boldsymbol{\lambda}^2 + \boldsymbol{\delta}^2)$
  - note $D(\boldsymbol{\lambda}\boldsymbol{\lambda}') = D(\lambda_1^2, \ldots, \lambda_J^2) = D(\boldsymbol{\lambda}^2)$

$$\boldsymbol{\Sigma} = D(\sqrt{\boldsymbol{\lambda}^2 + \boldsymbol{\delta}^2})\boldsymbol{\rho}D(\sqrt{\boldsymbol{\lambda}^2 + \boldsymbol{\delta}^2})$$

where

$$\boldsymbol{\rho} = \boldsymbol{\lambda}^*(\boldsymbol{\lambda}^*)' + (\boldsymbol{I} - D(\boldsymbol{\lambda}^*(\boldsymbol{\lambda}^*)'))$$

or

$$Z_{ij} = \frac{Y_{ij} - \mu_{y_j}}{\sigma_{y_j}}$$

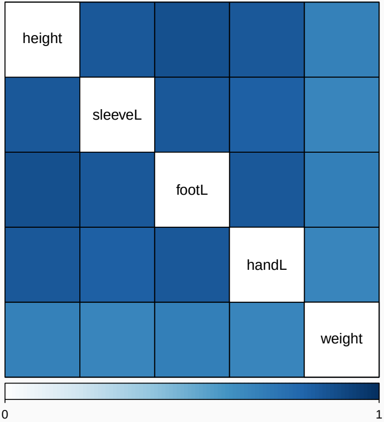
$$\boldsymbol{Z}_i \sim \mathrm{N}_J(\mathbf{0}, \boldsymbol{\rho})$$

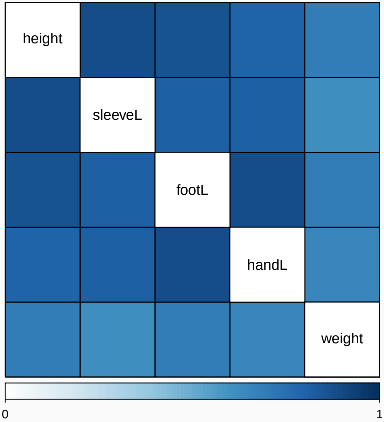
# Values

```r
inspect(fit[[1]],what="est")
```

```
## $lambda
##              f1
## height    8.889
## sleeveL   3.616
## footL     1.689
## handL     1.094
## weight   11.301
##
## $theta
##           height  sleevL   footL   handL  weight
## height    11.401
## sleeveL    0.000   2.945
## footL      0.000   0.000   0.389
## handL      0.000   0.000   0.000   0.267
```

# How Good Is The Fit?



# SEM Model Represenation

```r
library(lavaanPlot)
lavaanPlot(model=fit[[1]],covs=T,coefs=T)
```

# SEM Model Represenation

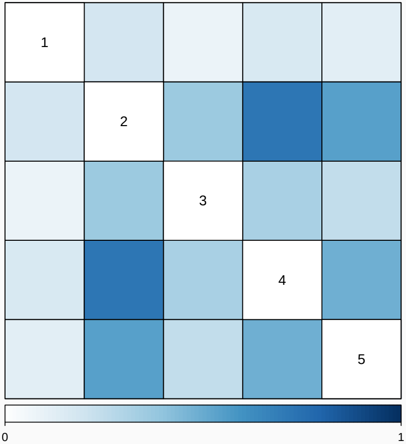
Jeff, Doesnt work for pdf!, go to rstudio

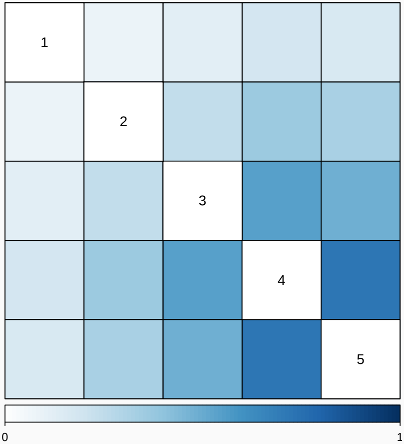
# Hypothetical One Factor Example

$$\lambda_j = (.2, .9, .4, .8, .6)$$

- use clustering to order rows and columns
- corrplot(cor,order='hclust')

# Hypothetical One Factor Example



# The General (Orthogonal) Factor Model

- Let $D$ be the number of factors
- Let $\phi_{di}$ be ability for the $i$th person on the $d$th factor
- Let $\lambda_{jd}$ be the loading from the $j$th score to the $d$th factor

Conditional/Univariate/unscaled

$$Y_{ij} \mid (\phi_{i1}, \ldots, \phi_{iD}) \sim \mathrm{N}\left(\mu_j + \sum_d \lambda_{jd}\phi_{id},\ \ \delta_j^2\right)$$

$$\phi_{id} \sim \mathrm{N}(0, 1)$$

or, for scaled,

$$Z_{ij} \mid (\phi_{i1}, \ldots, \phi_{iD}) \sim \mathrm{N}\left(\sum_d \lambda_{jd}\phi_{id},\ \ 1 - \sum_d \lambda_{jd}^2\right)$$

$$\phi_{id} \sim \mathrm{N}(0, 1)$$

# The Orthogonal Factor Model

Conditional + Multivariate

$$\boldsymbol{\lambda}_{J\times D} = \begin{pmatrix} \lambda_{11} & \dots & \lambda_{1D} \\ \vdots & \vdots & \vdots \\ \lambda_{J1} & \dots & \lambda_{JD} \end{pmatrix}$$

$$\begin{aligned} \boldsymbol{Z}_i|\phi_i &\sim \mathrm{N}_J(\boldsymbol{\lambda}\phi,\ \boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\lambda}\boldsymbol{\lambda}')) \\ \phi_i &\sim \mathrm{N}_D(\mathbf{0}, \mathbf{I}) \end{aligned}$$

# The Orthogonal Factor Model

Marginal + Multivariate

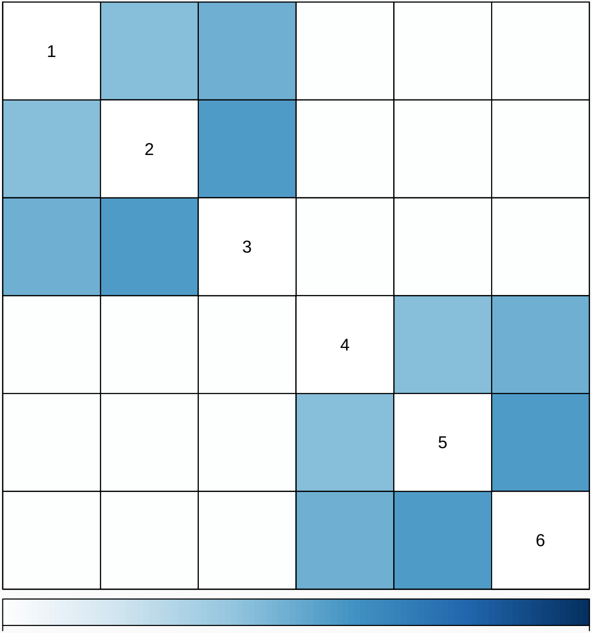
$$\boldsymbol{Z}_i \sim \mathrm{N}_J(\mathbf{0},\ \boldsymbol{\rho})$$
$$\boldsymbol{\rho} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\lambda}\boldsymbol{\lambda}')$$

# 2-Factor Example A

- 6 Scores/Tasks
- 2 Factors
- First 3 load on 1 factor; next three on another

$$\lambda_{\cdot 1} = (.6, .7, .8, 0, 0, 0)$$
$$\lambda_{\cdot 2} = (0, 0, 0, .6, .7, .8)$$

# 2-Factor Example A

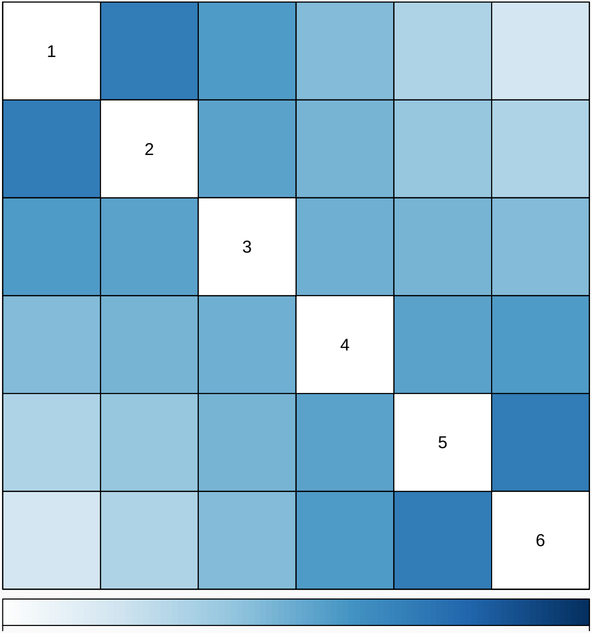


# 2-Factor Example B

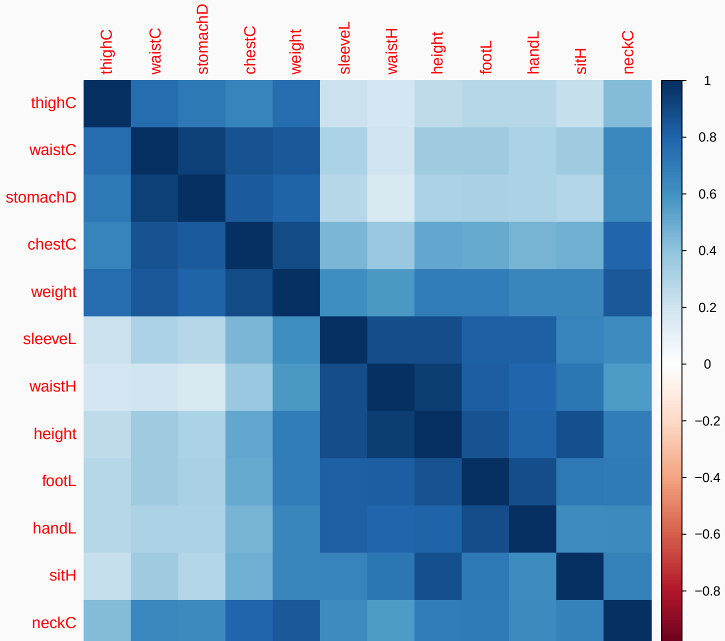
- 6 Scores/Tasks
- 2 Factors
- Graded Loadings

$$\lambda_{\cdot 1} = (.9, .74, .58, .42, .26, .1)$$
$$\lambda_{\cdot 2} = (.1, .26, .42, .58, .74, .9)$$

# 2-Factor Example B



# Anthropomorphic Example



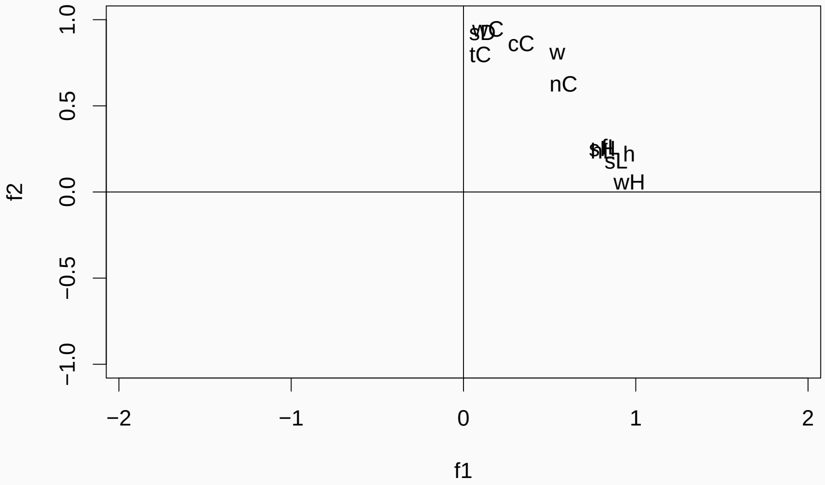
# Exploratory Factor Analysis (orthogonal)

```r
shortNames<-c('wH','fL','hL','sH','sL','nC','cC','wC','tC',
colnames(y) <- shortNames
z=scale(y)
z=as.data.frame.matrix(z)
fit <- efa(data=z,rotation='varimax',nfactors=1:5)
summary(fit)
```

```
## This is lavaan 0.6.17 -- running exploratory factor anal
##
##   Estimator                                          ML
##   Rotation method                      VARIMAX ORTHOGONAL
##   Rotation algorithm (rstarts)                   GPA (30)
##   Standardized metric                              TRUE
##   Row weights                                    Kaiser
##
##   Number of observations                            400
```

```
lavaanPlot(model = fit[[2]], coefs = T,covs=T)
```

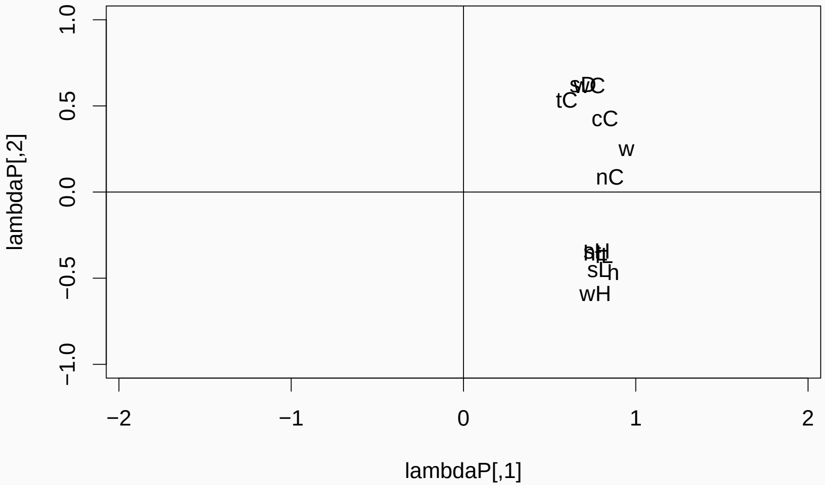
# Factor Loadings (Varimax)



# Rotation Issue

- The critical term in FA is $\boldsymbol{\lambda}\boldsymbol{\lambda}'$.
- Suppose there are two loadings $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ that are unique, that is $\boldsymbol{\lambda}_1 \neq \boldsymbol{\lambda}_2$. Yet, still, $\boldsymbol{\lambda}_1\boldsymbol{\lambda}_1' = \boldsymbol{\lambda}_2\boldsymbol{\lambda}_2'$
- Can it happen? Yes.
- A rotation matrix $\boldsymbol{A}$ has the following properties:
    - $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I} = \boldsymbol{A}\boldsymbol{A}'$, $det(\boldsymbol{A}) = \pm 1$
- $\boldsymbol{\lambda}_1\boldsymbol{\lambda}_1' = \boldsymbol{\lambda}_1\boldsymbol{I}\boldsymbol{\lambda}_1' = \boldsymbol{\lambda}_1\boldsymbol{A}\boldsymbol{A}'\boldsymbol{\lambda}_1' = (\boldsymbol{\lambda}_1\boldsymbol{A})(\boldsymbol{A}'\boldsymbol{\lambda}_1') = (\boldsymbol{\lambda}_1\boldsymbol{A})(\boldsymbol{\lambda}_1\boldsymbol{A})'$
- $\boldsymbol{\lambda}_2 = \boldsymbol{\lambda}_1\boldsymbol{A}$
- $\boldsymbol{\lambda}_1\boldsymbol{\lambda}_1' = \boldsymbol{\lambda}_2\boldsymbol{\lambda}_2'$

# PCA-Like Rotation



# Non-Orthogonal Factor Models

- Correlation among abilities across people.
- Allow correlation among latent people abilities, $\phi_i$.

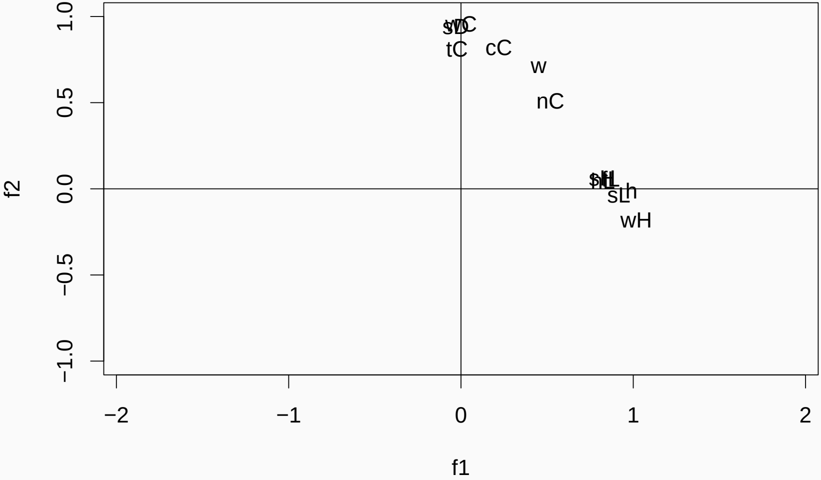
$$\boldsymbol{Z}_i|\phi_i \sim \mathrm{N}_J(\boldsymbol{\lambda}\phi,\ \boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\lambda}\boldsymbol{\lambda}'))$$
$$\phi_i \sim \mathrm{N}_D(\mathbf{0}, \psi)$$

where $\boldsymbol{\psi}$ is a correlation matrix describing relations among factors.

$$\boldsymbol{Z}_i \sim \mathrm{N}_J(\mathbf{0},\ \boldsymbol{\rho})$$
$$\boldsymbol{\rho} = \boldsymbol{\lambda}\psi\boldsymbol{\lambda}' + \boldsymbol{I} - \boldsymbol{D}(\boldsymbol{\lambda}\psi\boldsymbol{\lambda}')$$

```r
fit <- efa(data=z,nfactors=2)
summary(fit)
```

```
## This is lavaan 0.6.17 -- running exploratory factor anal
##
##   Estimator                                            ML
##   Rotation method                          GEOMIN OBLIQUE
##   Geomin epsilon                                    0.001
##   Rotation algorithm (rstarts)                    GPA (30)
##   Standardized metric                                TRUE
##   Row weights                                        None
##
##   Number of observations                              400
##
## Fit measures:
##                       aic      bic    sabic    chisq df pv
##   nfactors = 2 7112.153 7251.854 7140.797 1384.904 43
##
```



tC
cC
w
nC
sL
wH
f1
f2
−2
−1
0
1
2
−1.0
−0.5
0.0
0.5
1.0

# Confirmatory Factor Model For Anthropomorphic Set

- Zero out some loadings

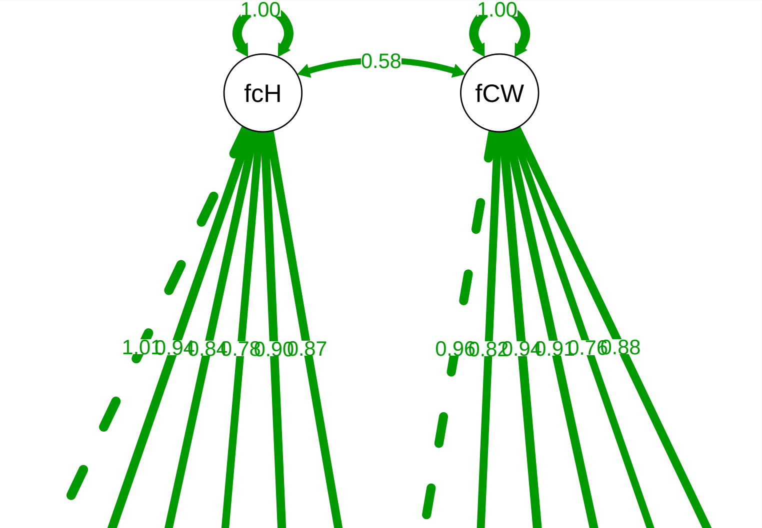
# Confirmatory Factor Model For Anthropomorphic Set

```r
library(semPlot)
model <- '
facH  =~ h+wH+fL+hL+sH+sL
facCW =~ w+nC+cC+wC+tC+sD'
fit=cfa(model,data=z)
summary(fit)
```

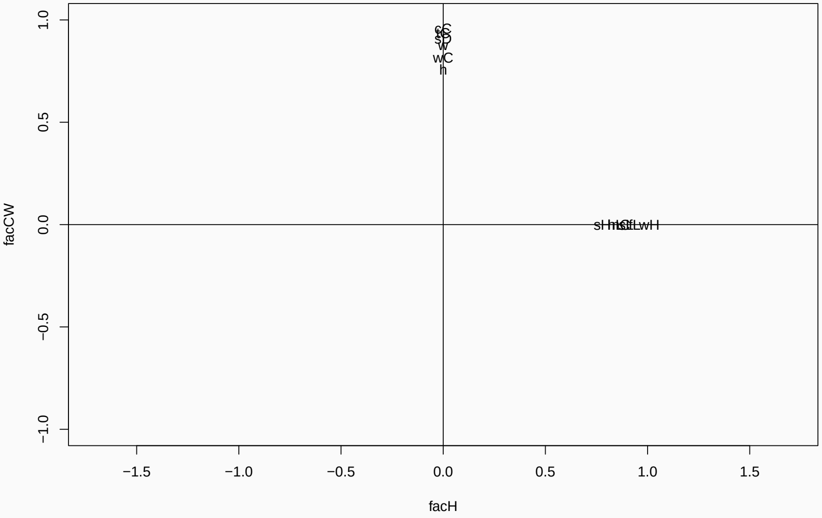
```
## lavaan 0.6.17 ended normally after 41 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                        25
##
##   Number of observations                           400
##
## Model Test User Model:
```

# Plots

```
semPaths(fit, "std", layout = "tree", intercepts = F, resic
         label.cex = 1, edge.label.cex=.95, fade = F)
```



# Full Separation!



# Your Turn

Illusions Data!

- 10 tasks (5 illusion types X 2 versions)
- 138 ppl
- score10.dat is a text file `read.table(...)`
- Please factor analyze using lavaan

# Your Turn

Illusions Data!

- suppose we are uninterested in version correlations?