Basic Data Analysis

Stephen Turnbull

Business Administration and Public Policy

Lecture 11: June 20, 2013

Abstract

Advanced multivariate methods and inference. A brief introduction to data mining as an alternative/complement to "classical" statistics.

Final Examination

- The final examination for this class will be held in **8A108** on Thursday, June 27 from 12:15–15:00.
- I plan to include content that was also on the midterm (about 1/3 and no more than 1/2 of the questions), as well as material covered since the midterm (at least 1/2). Conceptual material will be the majority as with the midterm.
- Length will be greater than the midterm, but not 2X as long.

Review Session

- A review session will be scheduled, probably on Friday, June 21, or Monday, June 24, from 5pm-7pm.
- Send mail to `data-vote@turnbull.sk.tsukuba.ac.jp` to expression your preference for date.
- The mail should have the following content:

  line 1: Your student ID

  line 2: Preferred date/time

  line 3: Two dashes and nothing else: `--`

4 and up: Any other comments about the review session.

- Mail is due by June 18, 09:00 (to allow preparation, reserving room *etc.*)

Structural modeling

- *Structural modeling* refers to a combination of a domain model and a statistical model that removes ambiguity from the interpretation of results. We say such models are *identified*.
- For example, because there is an infinite number of ways to define two lines that cross in a given point, it is impossible to reconstruct either a supply curve or a demand curve merely from price and quantity data.
- A regression line that relates price data to quantity data need have no relationship to either supply or demand.
- Addition of variables that only shift supply (*e.g.*, cost factors) allows identification of the demand curve.
- Addition of variables that only shift demand (*e.g.*, consumer income) allows identification of the supply curve.

- The statistical model is very simple: the supply and demand curves have disturbances that are additive, independent of each other, and i.i.d. over time.

Structural modeling

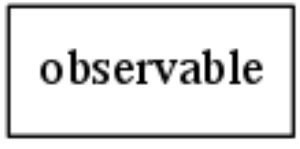
- The supply and demand model is a *simultaneous equation regression model*, where we have several equations, and the specific form of these equations (*i.e.*, whether certain variables are present in one or the other or both) determines whether the model is identified.
- In other cases we may be unwilling to specify equations. For example, our data may be qualitative, so that adding and multiplying by coefficients can't be justified. In these cases, we can use *confirmatory factor analysis* (CFA) and other kinds of *structural equation modeling* (SEM) to test whether our hypotheses that certain variables are related are supported by the data.
- Finally, we may be very uncertain about the domain model.

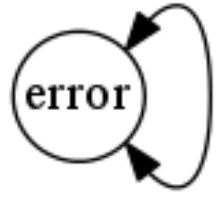
Here we can use *exploratory factor analysis* (EFA) to *discover* the important relationships in the data.

Graphical representation of structural models

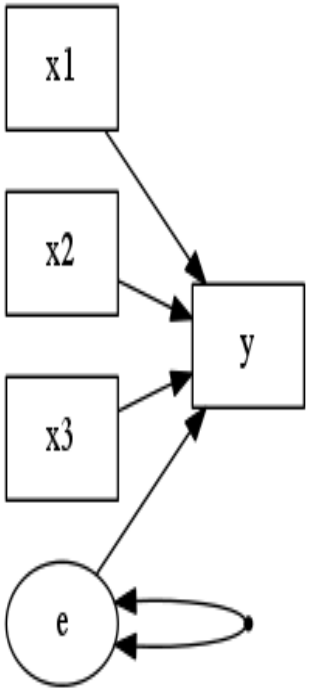
We can represent models (causal relationships) graphically.

- Rectangles indicate *observable* (explicit, manifest) variables.
- Ellipses indicate *latent* (unobservable) variables.
- One-headed arrows indicate *direct* (causal) *effects*, associated with equation *coefficients*.
- Two-headed arrows indicate *(co)variance*. These always connect exogenous variables.
- Errors are exogenous, latent variables.





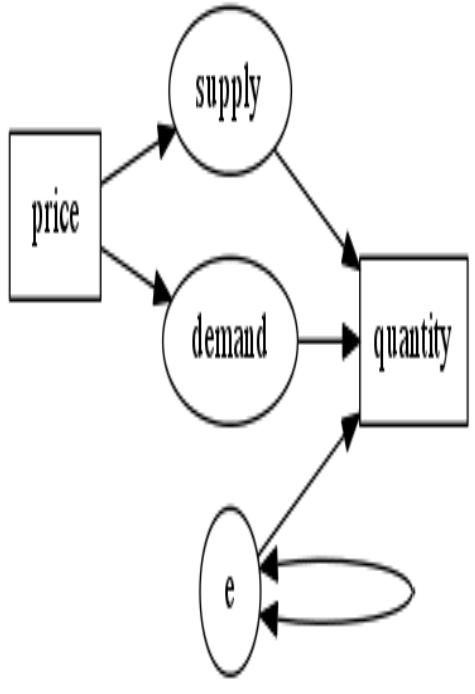
Graph for simple regression



- All explanatory variables are observable.
- Only the exogenous error e is latent.

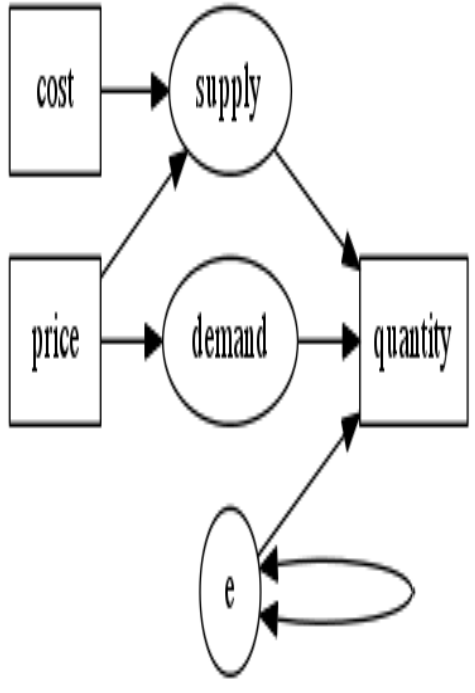
Graph for reduced form

- In economic theory, the relationship between price and quantity is mediated by two latent variables, supply and demand.
- In this graph, the equilibrium constraint (*quantity supplied* = *quantity demanded*) is not modeled.
- The latent variables are *underidentified*, that is, *not* identified. Their coefficients cannot be inferred from this model.
- Unlike the problem of multicollinearity, identification does *not* depend on actual data. It is a property of the theory and the variables available.



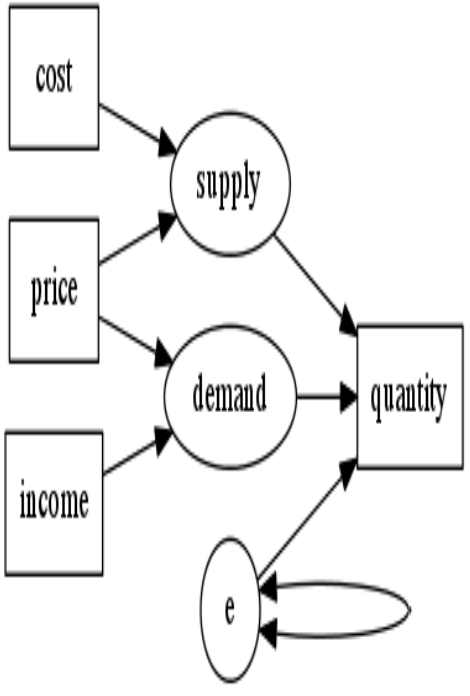
Instrumental variables

- By adding an *instrument*, a variable that affects one latent variable but not the other, we can identify the coefficient on the *second* variable. (The technical condition is that it not affect the *dependent* variable.)
- *Cost* should affect supply, but not demand (any effect of cost on demand is communicated to buyers through supply and price).
- *Demand* (coefficients) are identified.
- The full *model* remains *underidentified*. Some coefficients cannot be inferred from this model.



Full structure

- By adding a new instrument, (consumer) income, we can identify supply as well.
- Both latent variables are identified.
- We can now say the *model* is identified.
- **N.B.** *Income* is an instrument for *demand*, but it identifies *supply*. *Cost* is an instrument for *supply*, but it identifies *demand*.



Identification in complex models

- When there are many unidentified latent variables, or latent variables share instruments, the identification problem becomes more complex.
- The basic principle is the same: find *instruments* that affect some variables but are not correlated with the error term.
  - These can be used to *predict* those variables.
  - The predictions are *uncorrelated with the error*, so using the predicted variables instead of raw data satisfies the assumptions of the regression model.
  - As usual, this procedure will increase *reported* standard errors, but they will be unbiased. If you use OLS, the reported standard errors are likely to be *strongly biased downward*; you are *overestimating* the accuracy of the

regression.

Computing structural regression models

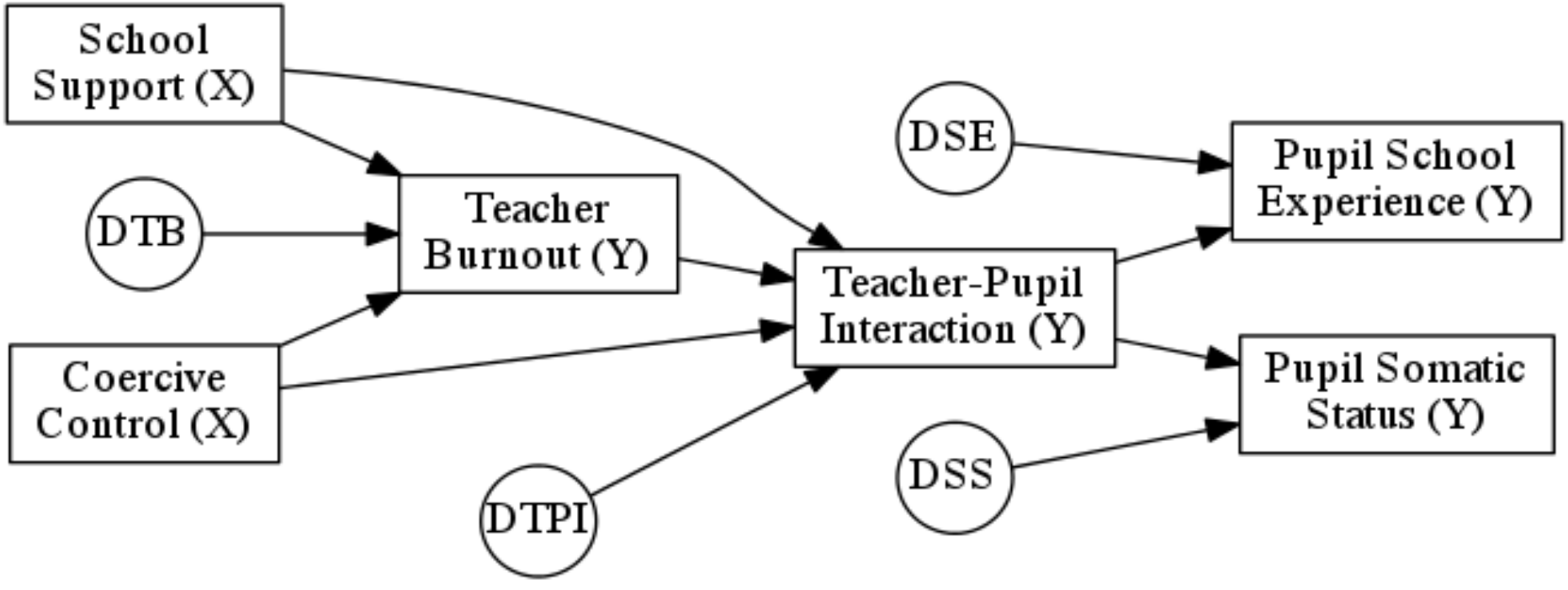
- In regression models, *two-stage least squares* (2SLS) can be used with instruments for linear models where some, but not all, latent relationships are identified.
  - Each model equation is estimated separately.
- For non-linear models, the *limited information maximum likelihood* (LIML) method is recommended.
- When the model is identified, the whole structure (all equations) can be recovered with *structural regression*, using *three-stage least squares* (3SLS) for linear models, and *full information maximum likelihood* (FIML) for nonlinear models.
  - The model equations are estimated *simultaneously*.

Structural equation modeling: SEM

- Economists and some other social scientists like *equilibrium* models, which are basically *fixed points* of a feedback loop (consider "scissors-paper-rock"), or solutions to *simultaneous equations*.
- Psychology and related fields often consider more complex structure. A set of techniques called *structural equation modeling* (SEM) has been developed to handle these.
- *Structural regression* is one of the techniques that is included in the field of SEM.

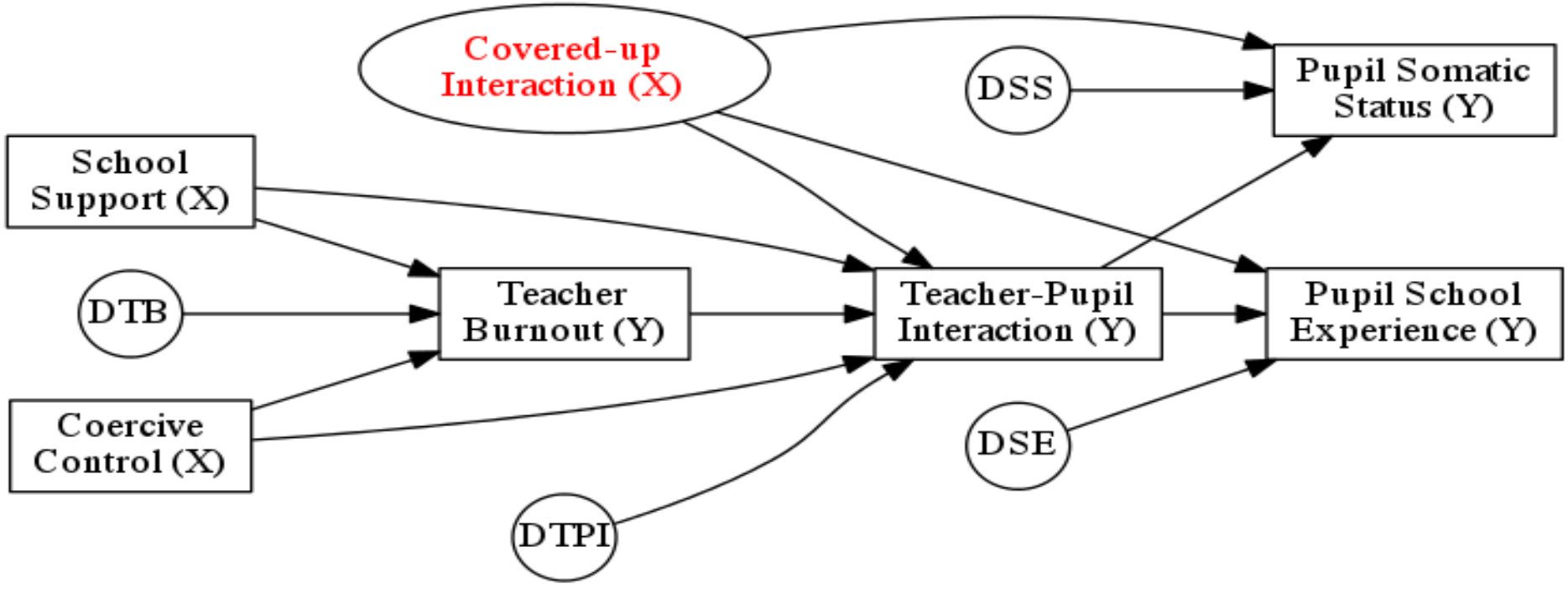
A complex model

Taken from Klein [2011], *Principles and Practice of Structural Equation Modeling*, 3e. (p. 111)



Issues in modeling in SEM

- This model assumes disturbances (the "D" circles) are uncorrelated with each other and with the endogenous variables.
- But there may be unmeasured aspects of teacher-pupil interaction, *e.g.*, due to cover-up of school problems. This causes *estimation bias*. The researcher must analyze such issues.



Resources for SEM

- Structural equation models are often called "LISREL models." *LISREL* was the first high-quality SEM software.
- Specialized SEM tools: Amos, LISREL, Mplus, and Mx. All have free downloads for students with limited functionality.
- Generic packages: SAS/STAT (CALIS module), SYSTAT (RAMONA module), and STATISTICA (SEPATH module).
- Check the documentation for your favorite statistics package; SEM is increasing in popularity.
- The best book is Rex B. Kline, *Principles and Practice of Structural Equation Modeling*, 3e, New York, Guilford Press, 2011. It has some of the best discussion of practical modeling issues I've ever seen, especially as in marketing and OB.

Factor Analysis

- In regression analysis, we assume we have a good idea explaining the behavior expressed in our data. We represent this explanation as a functional model.
  - Typically, a vector equation $y = f(x)$, *i.e.*,

$$\begin{aligned} y_1 &= f_1(x_1, \ldots, x_k) \\ &\vdots \\ y_n &= f_n(x_1, \ldots, x_k) \end{aligned}$$

  - Sometimes an implicit function: $0 = g(x, y)$.
- In factor analysis, we only have the dependent variables, y, and we want to find a small number of *factors* $x_1, \ldots, x_k$ that explain those variables.

A Simple Example

Consider the following data set, expressed in R:

```
v1 <- c(1,1,1,1,1,1,1,1,1,1,3,3,3,3,3,4,5,6)
v2 <- c(1,2,1,1,1,1,2,1,2,1,3,4,3,3,3,4,6,5)
v3 <- c(3,3,3,3,3,1,1,1,1,1,1,1,1,1,1,5,4,6)
v4 <- c(3,3,4,3,3,1,1,2,1,1,1,1,2,1,1,5,6,4)
v5 <- c(1,1,1,1,1,3,3,3,3,3,1,1,1,1,1,6,4,5)
v6 <- c(1,1,1,2,1,3,3,3,4,3,1,1,1,2,1,6,5,4)
```

- Ignoring the last three elements, `v1`, `v3`, and `v5` are data which are all 1s, except that the 3rd third, the 1st third, and the middle third, resp. are replaced by 3s.
- `v2`, `v4`, and `v6` are `v1`, `v3`, and `v5`, resp., with a little added "noise" (randomness).
- The last three elements ensure nonsingularity.

Correlations for the Simple Example

	v1	v2	v3	v4	v5	v6
v1	1.0000000	0.9393083	0.5128866	0.4320310	0.4664948	0.4086076
v2	0.9393083	1.0000000	0.4124441	0.4084281	0.4363925	0.4326113
v3	0.5128866	0.4124441	1.0000000	0.8770750	0.5128866	0.4320310
v4	0.4320310	0.4084281	0.8770750	1.0000000	0.4320310	0.4323259
v5	0.4664948	0.4363925	0.5128866	0.4320310	1.0000000	0.9473451
v6	0.4086076	0.4326113	0.4320310	0.4323259	0.9473451	1.0000000

- The correlations tell us how closely the variables are related to each other. It should not be surprising that `v1` and `v2` have a very high correlation, and so on.
- Similarly it should be plausible that `v1` and `v3` have a medium correlation.

What Do the Correlations Mean?

- These are artificial data, we know why they are correlated.
- "Eyeballing the numbers," or plotting them on a graph, also makes the relationship clear.
- Sometimes neither is true for "real data."
- We would like an automatic way to "extract" the "causes" of the measured behavior.
- *Factor analysis* of the correlations allows us to do this.

Can We Find Just One "Hidden Cause"?

We ask R to perform a one-factor analysis:

```
factanal(m1, factors = 1)
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.773 0.792 0.733 0.795 0.022 0.085


Loadings:
        v1     v2     v3     v4     v5     v6
Factor1 0.476  0.456  0.517  0.453  0.989  0.956

                Factor1
SS loadings       2.800
Proportion Var    0.467
```

```
Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 53.43 on 9 degrees of freedom.
The p-value is 2.43e-08
```

How About Two?

We ask R to perform a two-factor analysis:

```
factanal(m1, factors = 2)
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.114 0.642 0.742 0.005 0.097


Loadings:
        v1     v2     v3     v4     v5     v6
Factor1 0.971  0.917  0.429  0.363  0.254  0.205
Factor2 0.228  0.213  0.418  0.355  0.965  0.928


                Factor1 Factor2
SS loadings       2.206   2.190
Proportion Var    0.368   0.365
```

```
Cumulative Var   0.368   0.733

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 23.14 on 4 degrees of freedom.
The p-value is 0.000119
```

How About Three?

We ask R to perform a three-factor analysis:

```
factanal(m1, factors = 3)
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005


Loadings:
        v1    v2    v3    v4    v5    v6
Factor1  0.944    0.905    0.236    0.180    0.242    0.193
Factor2  0.182    0.235    0.210    0.242    0.881    0.959
Factor3  0.267    0.159    0.946    0.828    0.286    0.196


                Factor1 Factor2 Factor3
SS loadings       1.893   1.886   1.797
```

```
Proportion Var   0.316   0.314   0.300
Cumulative Var   0.316   0.630   0.929

The degrees of freedom for the model is 0 and the fit was 0.4755
```

Three with Rotation

We ask R to perform a three-factor analysis:

```
factanal(m1, factors = 3, rotation = "promax")
```

```
Uniquenesses:
   v1    v2    v3    v4    v5    v6
0.005 0.101 0.005 0.224 0.084 0.005


Loadings:

           v1      v2      v3      v4      v5      v6
Factor1                                   0.910   1.033
Factor2   0.985   0.951
Factor3                   1.003   0.867

               Factor1 Factor2 Factor3
```

```
SS loadings      1.903   1.876   1.772
Proportion Var   0.317   0.313   0.295
Cumulative Var   0.317   0.630   0.925

Factor Correlations:
        Factor1 Factor2 Factor3
Factor1   1.000  -0.462   0.460
Factor2  -0.462   1.000  -0.501
Factor3   0.460  -0.501   1.000

The degrees of freedom for the model is 0 and the fit was 0.4755
```

Why no test?

- You may have noticed that there was no report of a hypothesis test for the 3-factor model.
- The reason is that there are no degrees of freedom left (degrees of freedom were zero!)
- Calculating degrees of freedom for the factor analysis is complicated; leave it up to the program.

Is There Really an IQ?

R provides a number of sample datasets and programs, including one on measurements of intellectual ability. But is there a single factor ("IQ") that accounts for all intellectual performance?

```
factanal(factors = 1, covmat = ability.cov)
```

```
Loadings:
        general picture blocks   maze   reading  vocab
Factor1  0.682   0.384   0.502   0.300   0.877   0.849

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 75.18 on 9 degrees of freedom.
The p-value is 1.46e-12
```

It would appear not!

Multiple Factors in Ability

```
factanal(factors = 2, covmat = ability.cov, rotation = "promax")

Uniquenesses:
general picture  blocks    maze reading   vocab
  0.455   0.589   0.218   0.769   0.052   0.334

Loadings:
        general  picture  blocks    maze    reading   vocab
Factor1   0.364                               1.023    0.811
Factor2   0.470    0.671    0.932    0.508

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 6.11 on 4 degrees of freedom.
The p-value is 0.191
```

In this data set, it seems that there are just two different "kinds" of intelligence, which we could call "geometric" (or "visual") and "verbal". "General intelligence" is related to *both* factors.

What is data mining?

- Modern economic processes produce huge amounts of data.
- Detailed relationships are unclear. *E.g.*, serial correlation might be within a few minutes in the market for a given stock, or extend over years in the same case.
- Some phenomena are not understood at all.
- Use available data to discover them.

Data mining methods

- Simple examples: correlation analysis, stepwise regression.
- Principle component analysis: find the combinations of explanatory variables which best expresses all data.
  - Eigenvector, eigenvalue analysis of linear algebra.
- Lack of understanding of fundamental principles leads to *nonparametric* and even *distribution-free* analysis.
- Examples:
  - "Nearest-neighbor"
  - "Kernel smoothing"
  - "Local regression"

"Online" regression

- The regression equation (two variable with intercept) we used for the model $Y = \alpha + \beta X + \epsilon$ was

$$\begin{aligned} b_n &= \sum_{i=1}^{n} x_i y_i \Big/ \sum_{i=1}^{n} x_i^2 \\ a_n &= \bar{Y} - b_n \bar{X} \end{aligned}$$

  where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $x_i = X_i - \bar{X}$, and similarly for $\bar{Y}$ and y_i.

- If we are receiving new data frequently (*e.g.*, stock markets),
- It is often useful to change to an *updating* (online, recursive) algorithm.

An online regression

- By saving a small amount of information calculated time n, we can reduce calculation needed to produce new estimates when new data arrives. These variables are $\bar{X}_n = \sum_{i=1}^{n} X_i$, $\bar{Y}_n = \sum_{i=1}^{n} Y_i$, $D_n = \sum_{i=1}^{n} x_i^2$, and b_n.
- Calculating $\bar{X}_n$ from $\bar{X}_{n-1}$ is simple:

$$\begin{aligned} \bar{X}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\left(\sum_{i=1}^{n-1} X + X_n\right) = \frac{n-1}{n}\left(\frac{\sum_{i=1}^{n-1} X}{n-1} + \frac{1}{n}X_n\right) \\ &= \frac{n-1}{n}\bar{X}_{n-1} + \frac{1}{n}X_n \end{aligned}$$

An online regression, cont.

- Then

$$\begin{aligned}
\bar{Y}_n &= \frac{n-1}{n}\bar{Y}_{n-1} + \frac{1}{n}Y_n \\
b_n &= \left(\sum_{i=1}^{n-1} x_i y_i + x_n y_n\right) \Big/ \left(\sum_{i=1}^{n-1} x_i^2 + x_n^2\right) \\
&= \frac{D_{n-1}}{D_{n-1}x_i^2 + x_n^2} \frac{\sum_{i=1}^{n-1} x_i y_i + x_n y_n}{\sum_{i=1}^{n-1} x_i^2} \\
&= \frac{D_{n-1}}{D_{n-1} + x_n^2}\left(b_{n-1} + \frac{x_n y_n}{D_{n-1}}\right)
\end{aligned}$$

- and $a_n = \bar{Y}_n - b_n \bar{X}_n$ as usual.

Homework 8

Due Thursday, 2012-06-21, 11:45 am. Submit by email to `data-hw@turnbull.sk.tsukuba.ac.jp`. Your header should look like this:

```
From: a-student@sk.tsukuba.ac.jp
To: data-hw@turnbull.sk.tsukuba.ac.jp
Subject: Basic Data Analysis HW#8
```

The subject should be all half-width Roman letters (ASCII).

Get the data

Due June 20, 11:45.

1. Get the data set `Section1All_csv.csv` from the home page.

   This data set has several sections with different kinds of data. *After reading and thinking about the rest of the problems*, pick one section; using data across sections is a bad idea.

2. Input the data into your statistical package, and print out the data of the section (only!—no fair printing everything and editing the output) you have picked.

   There are two basic ways to accomplish this: create a new data set with exactly the rows and columns you need, or input the whole thing and use the package to pick out "your" variables. Also, many packages prefer that variables be columns and rows be observations, but this sheet has the opposite orientation.

Correlation matrix

3. Generate a correlation matrix for all the variables in your section.

4. Think of some way in which *some* of the variables in your section are related. Refer to scientific theory where possible.

Define and estimate a model

5. Define a regression model for *the variables you picked.*
   (a) Explain why you picked the dependent variable.
   (b) Write down your regression model.
   (c) Estimate the regression model using your statistical package.

Add an unrelated variable

6. Add a random, and therefore unrelated, variable to the model.
   (a) Use Excel or your statistical package to generate a series of random numbers, enough to make a new variable for your data set.
   (b) Add it to the data set, and print out the data set (*i.e.*, your model variables plus the random variable).
   (c) Add the random variable to your model of problem 5 as an explanatory variable, and estimate the new regression model.
   (d) Define and execute a hypothesis test that the new variable is in fact statistically unrelated to the model.

Homework 9

Due Thursday, 2011-06-21, 11:45 am. Submit by email to `data-hw@turnbull.sk.tsukuba.ac.jp`. Your header should look like this:

```
From: a-student@sk.tsukuba.ac.jp
To: data-hw@turnbull.sk.tsukuba.ac.jp
Subject: Basic Data Analysis HW#9
```

The subject should be all half-width Roman letters (ASCII).

Factor analysis of artificial data

1. Reproduce the factor analysis of six artificial variables done in class using your preferred statistical package.

Factor analysis of real data

2. Using the same data as in the regression problems, conduct a factor analysis on one factor, two factors, *etc.*, until you have "enough" factors.

3. Explain how you know when you have enough. Be quantitative!