Basic Data Analysis

Stephen Turnbull

Business Administration and Public Policy

Lecture 10: June 13, 2013

Abstract

Introduction to factor analysis and structural modeling.

An example session: Regression results

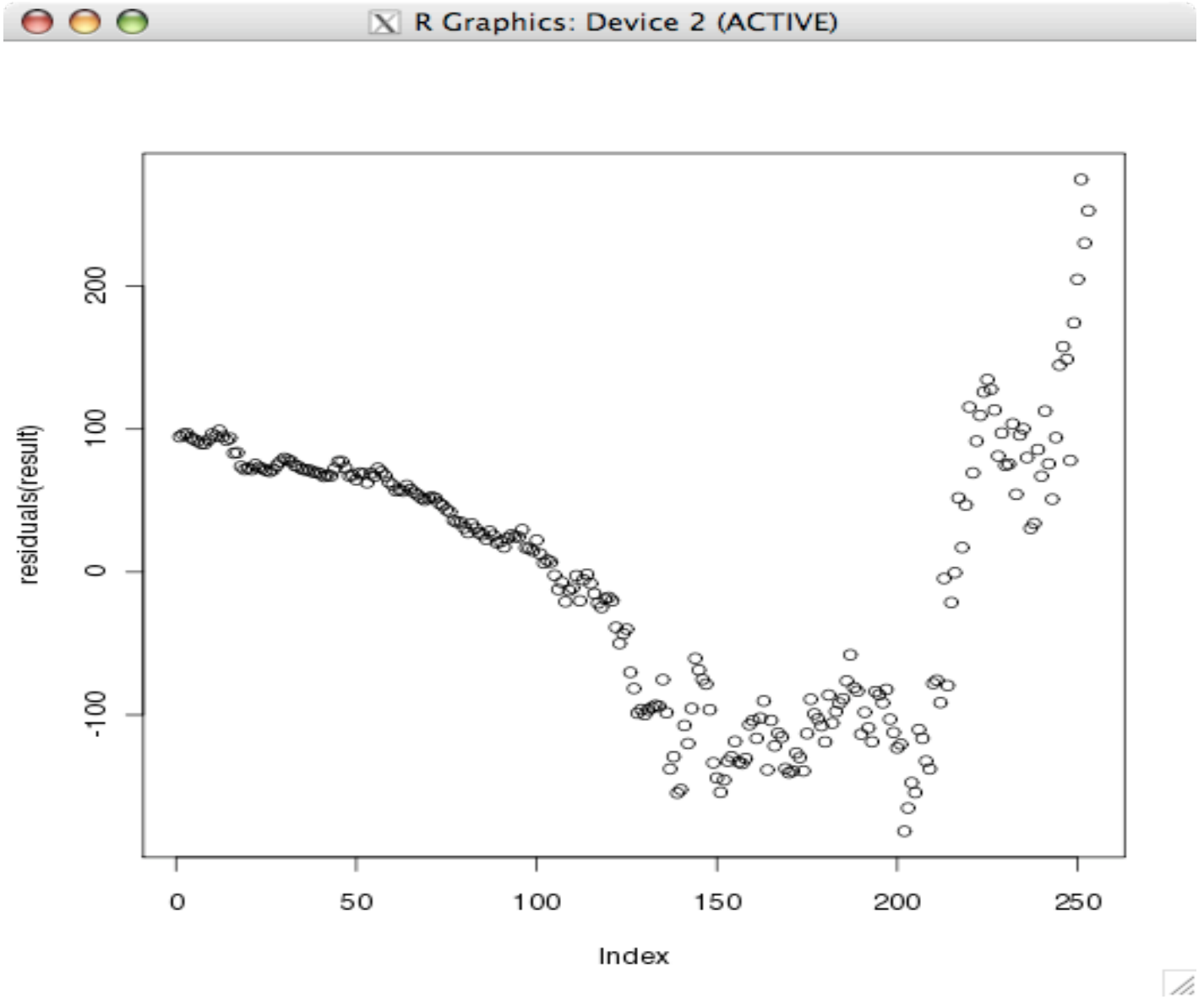
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.042e+02  8.056e+00  -12.93   <2e-16 ***
GDP          6.995e-01  1.321e-03  529.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.73 on 251 degrees of freedom
Multiple R-squared: 0.9991,     Adjusted R-squared: 0.9991
F-statistic: 2.802e+05 on 1 and 251 DF,  p-value: < 2.2e-16

> plot(residuals(result))
```

A simple graph

That doesn't look very random!



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.

Interpreting factor analysis

- A report of a factor analysis always contains the *factor loadings*.
- Report of the factor estimates themselves are optional (each factor has as many components as the data, and since they are not observable in reality, they are typically not very interesting).
- If v_i is an observed variable, and we have two factors f_1 and f_2, then v_i can be expressed

$$v_i = \hat{\alpha}_1 f_1 + \hat{\alpha}_2 f_2 + \epsilon_i.$$

- The *factor loadings* are the coefficients $\hat{\alpha}_j$.

Exploratory factor analysis

- The factor analyses we've discussed so far were *unrestricted*.
- The computations allowed *any* variable to load on *any* factor.
- Issues:
  - Theory may suggest some variables are *independent* of each other; then they should not have any factors in common.
  - Degrees of freedom are limited. We may have some idea of what factors there are, but an unrestricted analysis may pick up other factors.
- An unrestricted factor analysis is more commonly called an *exploratory factor analysis*.

Restrictions on factors

- Restrictions on factors are mathematically expressed as the statement that a factor loading is 0.
- These restrictions can derive from several theoretical considerations.
- One is *measurement*. We may measure several *indicators* of a single factor. These indicators should load on that factor, but no others.
- There may also be *structural* relations among the factors.

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.

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!

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.*)