Math 463, Spring 2009, University of Oregon

David A. Levin

University of Oregon

May 29, 2009

R: Crime rate: # of offenses reported to police per million population
Age: The number of males of age 14-24 per 1000 population
S: Indicator variable for Southern states (0 = No, 1 = Yes)
Ed: Mean # of years of schooling x 10 for persons of age 25 or older
Ex0: 1960 per capita expenditure on police by state and local government
Ex1: 1959 per capita expenditure on police by state and local government
LF: Labor force participation rate per 1000 civilian urban males age 14-24
M: The number of males per 1000 females
N: State population size in hundred thousands
NW: The number of non-whites per 1000 population
U1: Unemployment rate of urban males per 1000 of age 14-24
U2: Unemployment rate of urban males per 1000 of age 35-39
W: Median value of transferable goods and assets or family income in tens of dollars
X: The number of families per 1000 earning below 1/2 the median income

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-6.918e+02	1.559e+02	-4.438	9.56e-05	***
Age	1.040e+00	4.227e-01	2.460	0.01931	*
S	-8.308e+00	1.491e+01	-0.557	0.58117	
Ed	1.802e+00	6.496e-01	2.773	0.00906	**
Ex0	1.608e+00	1.059e+00	1.519	0.13836	
Ex1	-6.673e-01	1.149e+00	-0.581	0.56529	
LF	-4.103e-02	1.535e-01	-0.267	0.79087	
M	1.648e-01	2.099e-01	0.785	0.43806	
N	-4.128e-02	1.295e-01	-0.319	0.75196	
NW	7.175e-03	6.387e-02	0.112	0.91124	
U1	-6.017e-01	4.372e-01	-1.376	0.17798	
U2	1.792e+00	8.561e-01	2.093	0.04407	*
W	1.374e-01	1.058e-01	1.298	0.20332	
X	7.929e-01	2.351e-01	3.373	0.00191	**

Correlations

	R	Age	S	Ed	Ex0	Ex1	LF	M	N	NW	U1	U2	W	X
R	1.00	-0.09	-0.09	0.32	0.69	0.67	0.19	0.21	0.34	0.03	-0.05	0.18	0.44	-0.18
Age	-0.09	1.00	0.58	-0.53	-0.51	-0.51	-0.16	-0.03	-0.28	0.59	-0.22	-0.24	-0.67	0.64
S	-0.09	0.58	1.00	-0.70	-0.37	-0.38	-0.51	-0.31	-0.05	0.77	-0.17	0.07	-0.64	0.74
Ed	0.32	-0.53	-0.70	1.00	0.48	0.50	0.56	0.44	-0.02	-0.66	0.02	-0.22	0.74	-0.77
Ex0	0.69	-0.51	-0.37	0.48	1.00	0.99	0.12	0.03	0.53	-0.21	-0.04	0.19	0.79	-0.63
Ex1	0.67	-0.51	-0.38	0.50	0.99	1.00	0.11	0.02	0.51	-0.22	-0.05	0.17	0.79	-0.65
LF	0.19	-0.16	-0.51	0.56	0.12	0.11	1.00	0.51	-0.12	-0.34	-0.23	-0.42	0.29	-0.27
M	0.21	-0.03	-0.31	0.44	0.03	0.02	0.51	1.00	-0.41	-0.33	0.35	-0.02	0.18	-0.17
N	0.34	-0.28	-0.05	-0.02	0.53	0.51	-0.12	-0.41	1.00	0.10	-0.04	0.27	0.31	-0.13
NW	0.03	0.59	0.77	-0.66	-0.21	-0.22	-0.34	-0.33	0.10	1.00	-0.16	0.08	-0.59	0.68
U1	-0.05	-0.22	-0.17	0.02	-0.04	-0.05	-0.23	0.35	-0.04	-0.16	1.00	0.75	0.04	-0.06
U2	0.18	-0.24	0.07	-0.22	0.19	0.17	-0.42	-0.02	0.27	0.08	0.75	1.00	0.09	0.02
W	0.44	-0.67	-0.64	0.74	0.79	0.79	0.29	0.18	0.31	-0.59	0.04	0.09	1.00	-0.88
X	-0.18	0.64	0.74	-0.77	-0.63	-0.65	-0.27	-0.17	-0.13	0.68	-0.06	0.02	-0.88	1.00

Leaving out correlated variables yields

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -463.3329    93.7402  -4.943 1.35e-05 ***
Age            0.8579     0.3572   2.402 0.020931 *
Ed             1.6947     0.4590   3.692 0.000649 ***
Ex0            1.3214     0.1455   9.080 2.30e-11 ***
U1             0.1932     0.1897   1.018 0.314569
X              0.6452     0.1527   4.226 0.000130 ***
```

Colinearity

- (Approximate) *Colinearity* occurs when there are near linear relationships among the covariates.
- If one of the x-variables can be written as a linear combination of some of the others, then there exists exact colinearity. In this case, $X^T X$ is singular, and there is no unique least squares estimate of β.
- Let R_i^2 be the r-squared value when regressing x_i on the other x_j's. Values of R_i^2 near 1 indicate a problem.
- $$\text{Var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \frac{1}{S_{x_j x_j}},$$

  where $S_{x_j x_j} = \sum_i (x_{ij} - \bar{x}_j)^2$.
- $1/(1 - R_j^2)$ is called variance inflation factor.

- $R_{U1} = 0.8316$, $R_{Ex1} = 0.9899$ indicates problems with colinearity.
- The function `vif` in the `faraway` package computes all the variance inflation factos:

```
> vif(model.matrix(lm(R~.,data=crime)))
(Intercept)         Age           S          Ed         Ex0         Ex1          LF           M
   2.017100    2.698021    4.876751    5.049442   94.633118   98.637233    3.677557    3.658444
          N          NW          U1          U2           W           X
   2.324326    4.123274    5.938264    4.997617    9.968958    8.409449
```

What to do about colinearity?

Simple solution: drop some of the variables.

Generalized Least Squares (GLS)

- Suppose $\varepsilon \sim \sigma^2 \Sigma$.
- Write $\Sigma = SS^T$. If

$$Y = X\beta + \varepsilon\,,$$

then

$$\underbrace{S^{-1}Y}_{Y'} = \underbrace{S^{-1}X}_{X'}\beta + \underbrace{S^{-1}\varepsilon}_{\varepsilon'}\,.$$

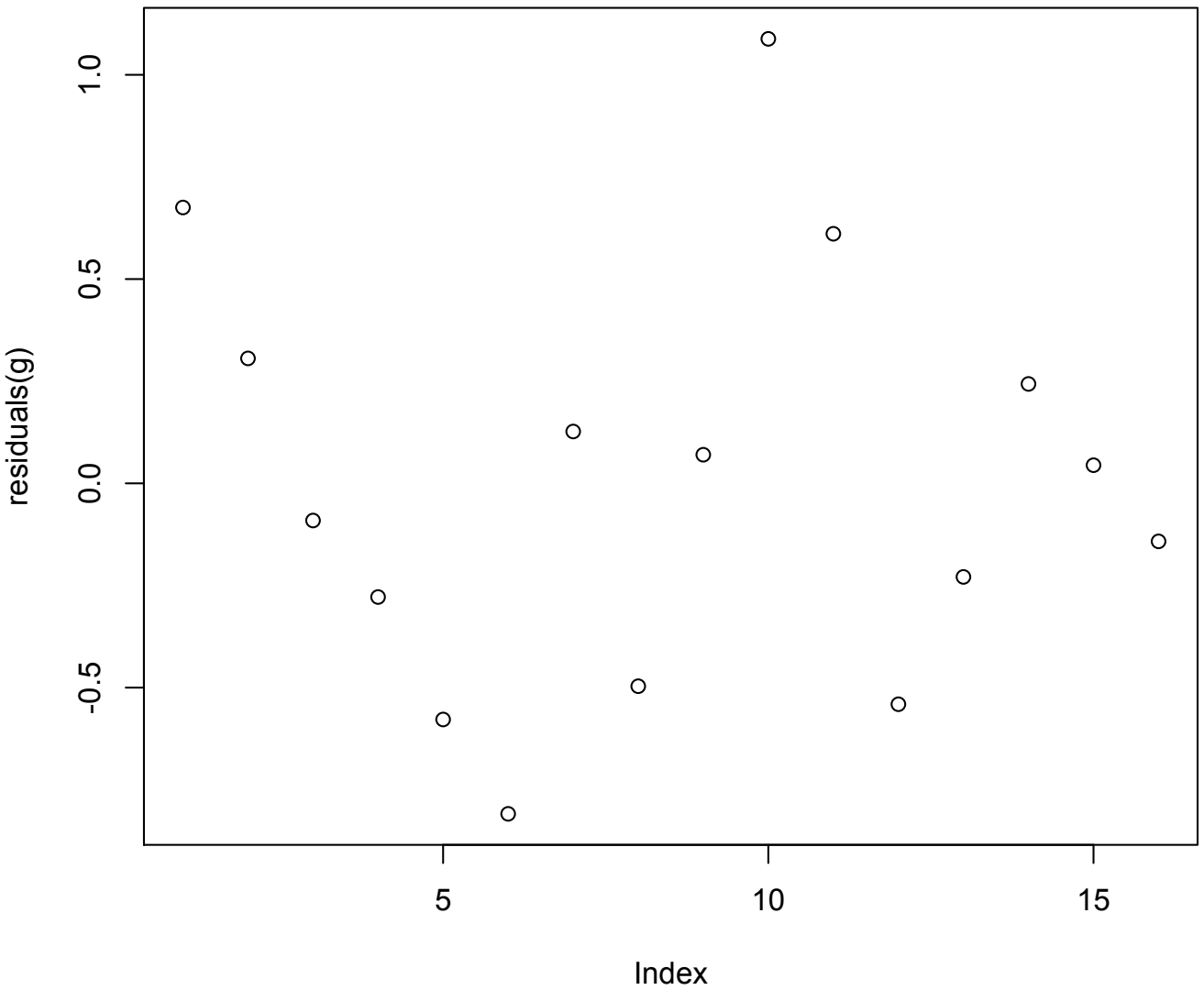
Then $\mathrm{Var}(Y') = \mathrm{Var}\,\varepsilon' = \sigma^2 S^{-1}\Sigma {S^{-1}}^T = \sigma^2 S^{-1} SS^T {S^{-1}}^T = \sigma^2 I.$

```
longley
Format

A data frame with 7 economical variables, observed yearly from 1947 to 1962 (n=16).

GNP.deflator:
GNP implicit price deflator (1954=100)
GNP:
Gross National Product.
Unemployed:
number of unemployed.
Armed.Forces:
number of people in the armed forces.
Population:
noninstitutionalized population >= 14 years of age.
Year:
the year (time).
Employed:
number of people employed.
```

Residuals vs. time



- Model the errors as

$$\varepsilon_{i+1} = \rho\varepsilon_i + \delta_{i+1}$$

  Then $\Sigma_{ij} = \rho^{|i-j|}$.
- Estimate ρ by correlation between

$$(\hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n)$$

  and

$$(\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_{n-1}).$$

  In this case, $\hat{\rho} = 0.310$.

```
> summary(lm(sy~sx-1))

Call:
lm(formula = sy ~ sx - 1)

Residuals:
       Min          1Q      Median          3Q         Max
-0.9851492 -0.3783524  0.0005565  0.2392798  0.8946370

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
sx(Intercept) 95.24972   13.81068   6.897 1.09e-05 ***
sxGNP          0.06777    0.01066   6.358 2.50e-05 ***
sxPopulation  -0.47859    0.15210  -3.147  0.00772 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5394 on 13 degrees of freedom
Multiple R-squared: 0.9999, Adjusted R-squared: 0.9999
F-statistic: 4.392e+04 on 3 and 13 DF,  p-value: < 2.2e-16
```

gls from nlme

```
> library(nlme)
> g <- gls(Employed~GNP+Population,correlation=corAR1(form=~Year), data=longley)
> summary(g)
Generalized least squares fit by REML
  Model: Employed ~ GNP + Population
  Data: longley
       AIC      BIC    logLik
  44.66377 47.48852 -17.33188

Correlation Structure: AR(1)
 Formula: ~Year
 Parameter estimate(s):
      Phi
0.6441692

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 101.85813 14.198932  7.173647  0.0000
GNP           0.07207  0.010606  6.795485  0.0000
Population   -0.54851  0.154130 -3.558778  0.0035

 Correlation:
           (Intr) GNP
GNP         0.943
Population -0.997 -0.966

Standardized residuals:
       Min         Q1        Med         Q3        Max
-1.5924564 -0.5447822 -0.1055401  0.3639202  1.3281898

Residual standard error: 0.689207
Degrees of freedom: 16 total; 13 residual
```

```
> intervals(g)
Approximate 95% confidence intervals

 Coefficients:
                  lower          est.       upper
(Intercept) 71.18320461 101.85813306 132.5330615
GNP          0.04915865   0.07207088   0.0949831
Population  -0.88149053  -0.54851350  -0.2155365
attr(,"label")
[1] "Coefficients:"

 Correlation structure:
         lower      est.    upper
Phi -0.4455798 0.6441692 0.964707
attr(,"label")
[1] "Correlation structure:"

 Residual standard error:
    lower      est.     upper
0.2472357 0.6892070 1.9212688
```

Weighted Least Squares

Suppose

$$\Sigma = \begin{bmatrix} 1/w_1 & 0 & \cdots & 0 \\ 0 & 1/w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/w_n \end{bmatrix}$$

Then

$$S^{-1} = \begin{bmatrix} \sqrt{w_1} & 0 & \cdots & 0 \\ 0 & \sqrt{w_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{w_n} \end{bmatrix}$$