

THE EFFECTS OF SPECIFICATION ERROR

The purpose of this handout is to illustrate the impact of specification error on parameter estimation in the multiple regression model. Assume there exists a dependent variable (Y) which is affected by three independent variables, X_1 through X_3 . Two of the independent variables are linearly related to each other: The bivariate (population) correlation between X_2 and X_3 is 0.55. Similarly, X_3 is linearly related to another variable, X_4 : The population correlation between X_3 and X_4 is also 0.55. The *true* (and usually unknown) population model relating the independent variable to the dependent variables is as follows:

$$Y = 10 + 0.3 X_1 + 0.5 X_2 + 0.4 X_3 + \epsilon \qquad R^2 = 0.75$$

One thousand observations are sampled randomly from this population. From this sample, univariate descriptive statistics and bivariate correlation coefficients are calculated, as shown below (in the R session listing). Then OLS is used to estimate the coefficients of several different regression equations. Note that the first equation shown is actually the “correct” one; that is, the specification that actually corresponds to the population structure. However, this information is usually unknown, so it is not unreasonable for a researcher to try the other equations shown below. Note the effects of the specification errors on the slope coefficients and the standard errors.

```
> library(RWinEdt)
>
> #####
> #####
> ###   The consequences of specification
> ###   error in regression models-- excluding
> ###   relevant variables and including
> ###   irrelevant variables
> #####
> #####
>
> ###   Read simulated data and name variables
>
> sim.data <- read.table(file.choose())
>
> colnames(sim.data) <- c("y", "x1", "x2", "x3", "x4")
>
> ###   Obtain summary statistics
>
> summary(sim.data)
      y              x1              x2              x3
Min.   : 1.00   Min.   : 0.05   Min.   : 0.07   Min.   : 0.47
1st Qu.: 52.93  1st Qu.:24.11  1st Qu.:26.14  1st Qu.:38.39
Median : 69.42  Median :49.84  Median :48.70  Median :49.28
Mean   : 69.87  Mean   :49.44  Mean   :50.72  Mean   :49.18
3rd Qu.: 87.14  3rd Qu.:73.93  3rd Qu.:76.70  3rd Qu.:59.99
Max.   :138.08  Max.   :99.83  Max.   :99.99  Max.   :99.63
      x4
Min.   : 0.01
1st Qu.:23.60
Median :50.32
Mean   :49.48
3rd Qu.:74.48
Max.   :99.92
>
> sd(sim.data)
      y              x1              x2              x3              x4
24.66249 28.69846 28.88889 16.85256 29.17553
>
```

The Effects of Specification Error

Page 2

```
> cor(sim.data)
      y          x1          x2          x3          x4
y  1.0000000  0.369142178  0.7508924333  0.602721866  0.1452876886
x1  0.3691422  1.000000000  -0.0097114141  0.005050134  0.0131740795
x2  0.7508924  -0.009711414  1.0000000000  0.544106805  -0.0008853677
x3  0.6027219  0.005050134  0.5441068048  1.000000000  0.5409076120
x4  0.1452877  0.013174080  -0.0008853677  0.540907612  1.0000000000
>
> ###
> ###   CASE 1: Estimate the "correct" model
> ###
>
> summary(lm(y ~ x1 + x2 + x3, data = sim.data))
```

```
Call:
lm(formula = y ~ x1 + x2 + x3, data = sim.data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-36.1113  -7.4109  -0.1464   7.9707  31.2544
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.23602     1.36117   6.051 2.04e-09 ***
x1           0.32112     0.01343  23.920 < 2e-16 ***
x2           0.51857     0.01590  32.624 < 2e-16 ***
x3           0.39560     0.02725  14.519 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 12.18 on 996 degrees of freedom
Multiple R-Squared:  0.757,    Adjusted R-squared:  0.7563
F-statistic: 1034 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
>
> ###
> ###   CASE 2: Estimate a model in which a correlated
> ###           independent variable, X2, is omitted
> ###           from the equation
>
> summary(lm(y ~ x1 + x3, data = sim.data))
```

```
Call:
lm(formula = y ~ x1 + x3, data = sim.data)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-57.0437 -11.3371  -0.3501  12.8056  53.0667
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.06644     1.95276   5.667 1.9e-08 ***
x1           0.31462     0.01930  16.304 < 2e-16 ***
x3           0.87933     0.03286  26.759 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 17.5 on 997 degrees of freedom
Multiple R-Squared:  0.4973,    Adjusted R-squared:  0.4963
F-statistic: 493.2 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
>
```

The Effects of Specification Error

Page 3

```
> ###
> ###   CASE 3: Estimate a model in which an uncorrelated
> ###       independent variable, X1, is omitted
> ###       from the equation
> ###
>
> summary(lm(y ~ x2 + x3, data = sim.data))

Call:
lm(formula = y ~ x2 + x3, data = sim.data)

Residuals:
    Min       1Q   Median       3Q      Max
-50.1324 -10.0835  -0.3763   9.9629  42.7296

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.00425    1.49355   16.07  <2e-16 ***
x2            0.51292    0.01993   25.73  <2e-16 ***
x3            0.40363    0.03417   11.81  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.27 on 997 degrees of freedom
Multiple R-Squared: 0.6174,    Adjusted R-squared: 0.6166
F-statistic: 804.4 on 2 and 997 DF,  p-value: < 2.2e-16

>
> ###
> ###   CASE 4: Estimate a model in which an irrelevant
> ###       independent variable, X4, is included
> ###       in the equation
> ###
> ###
>
> summary(lm(y ~ x1 + x2 + x3 + x4, data = sim.data))

Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = sim.data)

Residuals:
    Min       1Q   Median       3Q      Max
-36.419  -7.381  -0.211   7.977  31.405

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.276199    1.364699   6.064 1.88e-09 ***
x1            0.321165    0.013431  23.913 < 2e-16 ***
x2            0.515307    0.017506  29.436 < 2e-16 ***
x3            0.405856    0.035678  11.376 < 2e-16 ***
x4           -0.007703    0.017291  -0.445  0.656
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.18 on 995 degrees of freedom
Multiple R-Squared: 0.757,    Adjusted R-squared: 0.7561
F-statistic: 775.1 on 4 and 995 DF,  p-value: < 2.2e-16

>
>
```