Multiple Regression: Including spurious variables – A small example

Simulation I

Data generating model: $n = 1000$,

$$Y_i = 2 + 3x_{i1} - 2x_{i2} + 2x_{i3} + \epsilon_i$$

i.e. $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$, with $\sigma = 4$.

All predictors are influential in the model.

Simulation I: Fit Model $X_1 + X_2 + X_3$

```
> fit.global<-lm(Y~x1+x2+x3); summary(fit.global)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.04262    0.02154   94.81   <2e-16 ***
x1           2.95810    0.03334   88.71   <2e-16 ***
x2          -1.98782    0.02976  -66.79   <2e-16 ***
x3           1.94419    0.02654   73.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2024 on 996 degrees of freedom
Multiple R-squared:  0.8971,    Adjusted R-squared:  0.8968
F-statistic:  2895 on 3 and 996 DF,  p-value: < 2.2e-16

> anova(fit.global)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x1          1  62.668  62.668  1530.1 < 2.2e-16 ***
x2          1  73.124  73.124  1785.4 < 2.2e-16 ***
x3          1 219.866 219.866  5368.3 < 2.2e-16 ***
Residuals 996  40.793   0.041
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simulation I: Fit Model $X_1 + X_2$

```
> fit.12<-lm(Y~x1+x2);summary(fit.12)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.28134    0.03374   97.25   <2e-16 ***
x1           1.54940    0.06883   22.51   <2e-16 ***
x2          -1.16460    0.06964  -16.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5113 on 997 degrees of freedom
Multiple R-squared:  0.3425,    Adjusted R-squared:  0.3412
F-statistic: 259.7 on 2 and 997 DF,  p-value: < 2.2e-16

> anova(fit.12)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x1          1  62.668  62.668  239.70 < 2.2e-16 ***
x2          1  73.124  73.124  279.69 < 2.2e-16 ***
Residuals 997 260.658   0.261
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparison

- When the **correct** model $X_1 + X_2 + X_3$ is fitted, inference proceeds and produces correct estimates. The ANOVA table (line 30) reveals that

$$\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2, \beta_3) = 40.793 \qquad \mathrm{MS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2, \beta_3) = 0.041$$

- When the **incorrect** model $X_1 + X_2$ – which omits X_3 – is fitted, the estimates are incorrect. The ANOVA table (line 46) reveals that

$$\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2) = 260.658 \qquad \mathrm{MS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2) = 0.261$$

  and the partial F-statistics assessing the influence of X_1 and X_2 are much smaller (compare lines 21–22 and 44–45).

Comparison (cont.)

Thus, the denominator in the F-test statistic has increased. This is because of the fact that

$$\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2) = \mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2, \beta_3) + \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_3 | \beta_0, \beta_1, \beta_2)$$

– see lines 22, 23 and 30; we have

$$\begin{aligned} \mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2, \beta_3) &= 40.793 \\ \overline{\mathrm{SS}}_{\mathrm{R}}(\beta_3 | \beta_0, \beta_1, \beta_2) &= 219.866 \\ \mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2) &= 260.658 \end{aligned}$$

All of the $\overline{\mathrm{SS}}_{\mathrm{R}}(\beta_3 | \beta_0, \beta_1, \beta_2)$ sum of squares from the first fit has been passed into $\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2)$.

Therefore, we conclude that if X_3 IS influential, it must be included in the analysis that studies the influence of X_1 and X_2 using partial F-tests.

This is because omitting X_3 artificially inflates the sum of squares for the residuals.

Simulation II

Data generating model: $n = 1000$,

$$Y_i = 2 + 3x_{i1} - 2x_{i2} + \epsilon_i$$

i.e. $\beta = (\beta_0, \beta_1, \beta_2, 0)^\top$, with $\sigma = 4$.

Only X_1 and X_2 are influential in the model.

Simulation II: Fit Model $X_1 + X_2 + X_3$

```
> fit.global<-lm(Y~x1+x2+x3);summary(fit.global)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.97752    0.02089  94.685   <2e-16 ***
x1           2.99318    0.03232  92.599   <2e-16 ***
x2          -1.97675    0.02885 -68.511   <2e-16 ***
x3           0.02694    0.02572   1.047    0.295
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1962 on 996 degrees of freedom
Multiple R-squared:  0.9273,    Adjusted R-squared:  0.9271
F-statistic:  4235 on 3 and 996 DF,  p-value: < 2.2e-16

> anova(fit.global)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq   F value Pr(>F)
x1          1 280.659 280.659 7292.1642 <2e-16 ***
x2          1 208.250 208.250 5410.8046 <2e-16 ***
x3          1   0.042   0.042    1.0966 0.2953
Residuals 996  38.334   0.038
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simulation II: Fit Model $X_1 + X_2$

```
> fit.12<-lm(Y~x1+x2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.99468    0.01295  154.07   <2e-16 ***
x1           2.97366    0.02641  112.59   <2e-16 ***
x2          -1.96534    0.02672  -73.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1962 on 997 degrees of freedom
Multiple R-squared:  0.9272,    Adjusted R-squared:  0.9271
F-statistic:  6351 on 2 and 997 DF,  p-value: < 2.2e-16

> anova(fit.12)
Analysis of Variance Table

Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x1          1 280.659 280.659  7291.5 < 2.2e-16 ***
x2          1 208.250 208.250  5410.3 < 2.2e-16 ***
Residuals 997  38.376   0.038
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparison

- When the model $X_1 + X_2 + X_3$ – with a spurious predictor – is fitted, inference proceeds and produces correct estimates for all parameters; the parameter estimate for β_3 is near zero. The ANOVA table (line 71) reveals that

$$\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2, \beta_3) = 38.334 \qquad \mathrm{MS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2, \beta_3) = 0.038$$

- When the **correct** model $X_1 + X_2$ is fitted, the estimates are correct. The ANOVA table (line 46) reveals that

$$\mathrm{SS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2) = 38.376 \qquad \mathrm{MS}_{\mathrm{Res}}(\beta_0, \beta_1, \beta_2) = 0.038$$

Comparison (cont.)

Thus, the F statistics assessing the influence of X_1 and X_2 **do not change greatly** – see lines 68, 69, 92 and 93.

The key point is that, from line 70,

$$\overline{\mathrm{SS}}_{\mathrm{R}}(\beta_3|\beta_0, \beta_1, \beta_2) = 0.042$$

Comparison (cont.)

Therefore, we conclude that if X_3 IS NOT influential, including it in the analysis that studies the influence of X_1 and X_2 using partial F-tests does NOT compromise the results to any great degree.

The only downside from including X_3 is that it uses up one degree of freedom; an extra parameter, β_3, is being estimated, when in fact that parameter is zero in the data generating model.

When n is moderate to large, this has a negligible effect; when n is small, the effect may be more noticeable.