BS2280 - Econometrics 1

Lecture 11 - Part 2: Identifying Nonlinearities and Multicollinearity

Dr. Yichen Zhu

Structure of today's lecture

Intended Learning Outcomes

- Understanding what collinearity means
- Understanding the consequences of perfect collinearity and multicollinearity
- Detecting multicollinearity
- Mitigating the problems of multicollinearity

Motivation

$$wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$

- Do you suspect a higher correlation between *exper* and $exper^2$ within this model?
- Will this higher correlation between *exper* and $exper^2$ affect the estimations?

Perfect Collinearity: Background

- Remember the 6 OLS assumptions for the **multiple regression model**
  1. Assumption 1: The model is linear in parameters and correctly specified
  2. Assumption 2: There is no exact linear relationship amongst the X variables in the sample
  3. Assumption 3: The disturbance term has zero expectation
  4. Assumption 4: The disturbance term is homoscedastic
  5. Assumption 5: The values of the disturbance term have independent distributions
  6. Assumption 6: The disturbance term has a normal distribution
- If all these assumptions hold, the OLS estimates will have certain desirable properties (Best Linear Unbiased Estimator, BLUE)

Perfect Collinearity: Background

- Remember the 6 OLS assumptions for the **multiple regression model**
  1. Assumption 1: The model is linear in parameters and correctly specified
  2. Assumption 2: There is no exact linear relationship amongst the X variables in the sample
  3. Assumption 3: The disturbance term has zero expectation
  4. Assumption 4: The disturbance term is homoscedastic
  5. Assumption 5: The values of the disturbance term have independent distributions
  6. Assumption 6: The disturbance term has a normal distribution
- If all these assumptions hold, the OLS estimates will have certain desirable properties (Best Linear Unbiased Estimator, BLUE)

Perfect Collinearity: Background

- Remember the 6 OLS assumptions for the **multiple regression model**
  1. Assumption 1: The model is linear in parameters and correctly specified
  2. Assumption 2: There is no exact linear relationship amongst the X variables in the sample
  3. Assumption 3: The disturbance term has zero expectation
  4. Assumption 4: The disturbance term is homoscedastic
  5. Assumption 5: The values of the disturbance term have independent distributions
  6. Assumption 6: The disturbance term has a normal distribution
- If all these assumptions hold, the OLS estimates will have certain desirable properties (Best Linear Unbiased Estimator, BLUE)

Perfect Collinearity: Background

- Remember the 6 OLS assumptions for the **multiple regression model**
  1. Assumption 1: The model is linear in parameters and correctly specified
  2. Assumption 2: There is no exact linear relationship amongst the X variables in the sample
  3. Assumption 3: The disturbance term has zero expectation
  4. Assumption 4: The disturbance term is homoscedastic
  5. Assumption 5: The values of the disturbance term have independent distributions
  6. Assumption 6: The disturbance term has a normal distribution
- If all these assumptions hold, the OLS estimates will have certain desirable properties (Best Linear Unbiased Estimator, BLUE)

Perfect Collinearity: Background

- Remember the 6 OLS assumptions for the **multiple regression model**
  1. Assumption 1: The model is linear in parameters and correctly specified
  2. Assumption 2: There is no exact linear relationship amongst the X variables in the sample
  3. Assumption 3: The disturbance term has zero expectation
  4. Assumption 4: The disturbance term is homoscedastic
  5. Assumption 5: The values of the disturbance term have independent distributions
  6. Assumption 6: The disturbance term has a normal distribution
- If all these assumptions hold, the OLS estimates will have certain desirable properties (Best Linear Unbiased Estimator, BLUE)

Perfect Collinearity: Background

- Remember the 6 OLS assumptions for the **multiple regression model**
  1. Assumption 1: The model is linear in parameters and correctly specified
  2. Assumption 2: There is no exact linear relationship amongst the X variables in the sample
  3. Assumption 3: The disturbance term has zero expectation
  4. Assumption 4: The disturbance term is homoscedastic
  5. Assumption 5: The values of the disturbance term have independent distributions
  6. Assumption 6: The disturbance term has a normal distribution
- If all these assumptions hold, the OLS estimates will have certain desirable properties (Best Linear Unbiased Estimator, BLUE)

Perfect Collinearity: Background

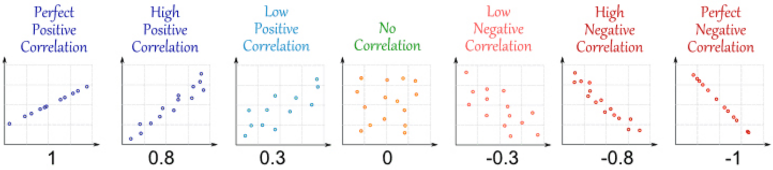
- Remember the 6 OLS assumptions for the **multiple regression model**
  1. Assumption 1: The model is linear in parameters and correctly specified
  2. Assumption 2: There is no exact linear relationship amongst the X variables in the sample
  3. Assumption 3: The disturbance term has zero expectation
  4. Assumption 4: The disturbance term is homoscedastic
  5. Assumption 5: The values of the disturbance term have independent distributions
  6. Assumption 6: The disturbance term has a normal distribution
- If all these assumptions hold, the OLS estimates will have certain desirable properties (Best Linear Unbiased Estimator, BLUE)

Perfect Collinearity: Background

- Remember the 6 OLS assumptions for the **multiple regression model**
  1. Assumption 1: The model is linear in parameters and correctly specified
  2. Assumption 2: There is no exact linear relationship amongst the X variables in the sample
  3. Assumption 3: The disturbance term has zero expectation
  4. Assumption 4: The disturbance term is homoscedastic
  5. Assumption 5: The values of the disturbance term have independent distributions
  6. Assumption 6: The disturbance term has a normal distribution
- If all these assumptions hold, the OLS estimates will have certain desirable properties (Best Linear Unbiased Estimator, BLUE)

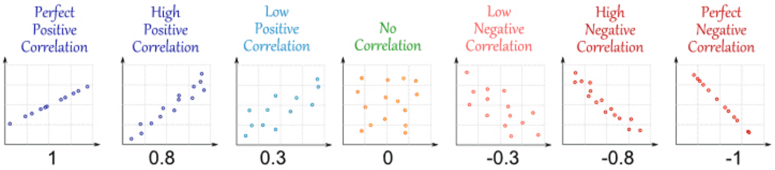
Perfect Collinearity: Background

- **Question**: What will happen if assumption 2 (There is no exact linear relationship amongst the X variables in the sample) does not hold?
- Actually, Assumption 2 states that there is no perfect collinearity between the X variables
- Perfect collinearity means that a X variable can be perfectly predicted linearly by other X variables
- If we calculate a correlation coefficient between X_1 and X_2, this correlation coefficient will be exactly 1 or -1



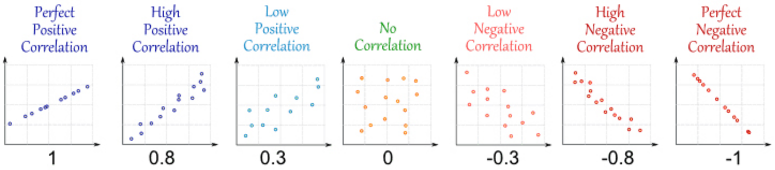
Perfect Collinearity: Background

- **Question**: What will happen if assumption 2 (There is no exact linear relationship amongst the X variables in the sample) does not hold?
- Actually, Assumption 2 states that there is no perfect collinearity between the X variables
- Perfect collinearity means that a X variable can be perfectly predicted linearly by other X variables
- If we calculate a correlation coefficient between X_1 and X_2, this correlation coefficient will be exactly 1 or -1



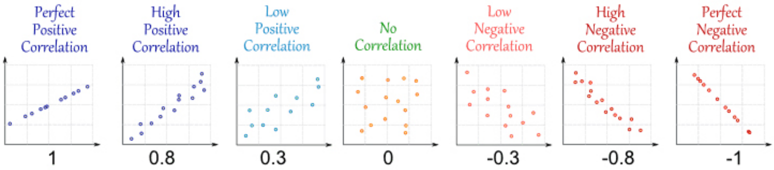
Perfect Collinearity: Background

- **Question**: What will happen if assumption 2 (There is no exact linear relationship amongst the X variables in the sample) does not hold?
- Actually, Assumption 2 states that there is no perfect collinearity between the X variables
- Perfect collinearity means that a X variable can be perfectly predicted linearly by other X variables
- If we calculate a correlation coefficient between X_1 and X_2, this correlation coefficient will be exactly 1 or -1



Perfect Collinearity: Background

- **Question**: What will happen if assumption 2 (There is no exact linear relationship amongst the X variables in the sample) does not hold?
- Actually, Assumption 2 states that there is no perfect collinearity between the X variables
- Perfect collinearity means that a X variable can be perfectly predicted linearly by other X variables
- If we calculate a correlation coefficient between X_1 and X_2, this correlation coefficient will be exactly 1 or -1



Perfect Collinearity: Example

- Consider two variables measuring age in days and age in weeks.

$$EARNINGS_i = \beta_1 + \beta_2 \textit{age in days} + \beta_3 \textit{age in weeks} + u_i$$

- Therefore, there is a perfect correlation or linear relationship between *age in days* and *age in weeks*.
- If you increase age in weeks by one unit, age in days will always increase by 7 units!!!

$$\textit{age in days} = 7 \textit{age in weeks}$$

- Then we will have some problems in estimating β_2 and β_3

Perfect Collinearity: Consequences

- If in our regression model are X variables that are perfectly collinear, then the software will either refuse to run the regression or it will drop one of the problematic X variable
- In practice, we will very rarely encounter perfect collinearity
- Perfect collinearity is easy to spot!
- We can then exclude these variables from the regression model

Perfect Collinearity: Consequences

- If in our regression model are X variables that are perfectly collinear, then the software will either refuse to run the regression or it will drop one of the problematic X variable
- In practice, we will very rarely encounter perfect collinearity
- Perfect collinearity is easy to spot!
- We can then exclude these variables from the regression model

Perfect Collinearity: Consequences

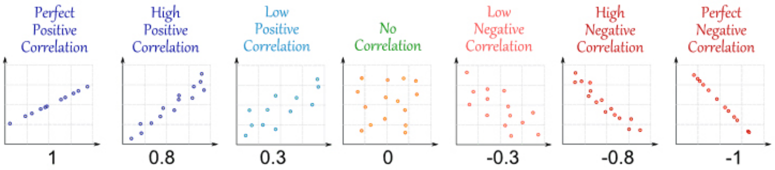
- If in our regression model are X variables that are perfectly collinear, then the software will either refuse to run the regression or it will drop one of the problematic X variable
- In practice, we will very rarely encounter perfect collinearity
- Perfect collinearity is easy to spot!
- We can then exclude these variables from the regression model

Perfect Collinearity: Consequences

- If in our regression model are X variables that are perfectly collinear, then the software will either refuse to run the regression or it will drop one of the problematic X variable
- In practice, we will very rarely encounter perfect collinearity
- Perfect collinearity is easy to spot!
- We can then exclude these variables from the regression model

Multicollinearity: Defination

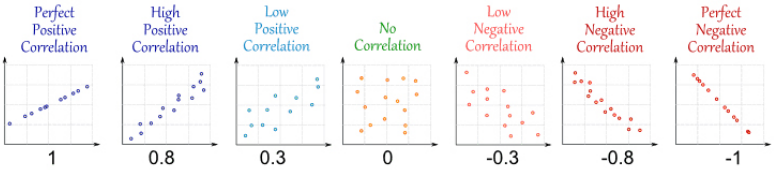
- Less than perfect collinearity is a more common occurrence than perfect collinearity
- This is a case when correlation exists between *X* variables that move together, but that correlation is not perfect
- The correlation coefficient will be between -1 and 1 but will never be exactly -1 or 1



- This problem of less than perfect collinearity is known as **multicollinearity**

Multicollinearity: Defination

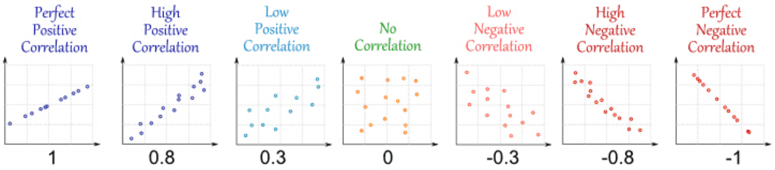
- Less than perfect collinearity is a more common occurrence than perfect collinearity
- This is a case when correlation exists between *X* variables that move together, but that correlation is not perfect
- The correlation coefficient will be between -1 and 1 but will never be exactly -1 or 1



- This problem of less than perfect collinearity is known as **multicollinearity**

Multicollinearity: Defination

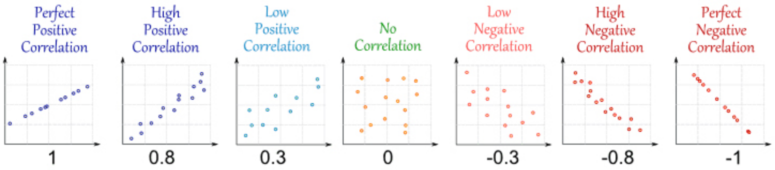
- Less than perfect collinearity is a more common occurrence than perfect collinearity
- This is a case when correlation exists between *X* variables that move together, but that correlation is not perfect
- The correlation coefficient will be between -1 and 1 but will never be exactly -1 or 1



- This problem of less than perfect collinearity is known as **multicollinearity**

Multicollinearity: Defination

- Less than perfect collinearity is a more common occurrence than perfect collinearity
- This is a case when correlation exists between X variables that move together, but that correlation is not perfect
- The correlation coefficient will be between -1 and 1 but will never be exactly -1 or 1



- This problem of less than perfect collinearity is known as **multicollinearity**

Multicollinearity: Example

- Multicollinearity is a problem in regression models where the X variables are correlated with each other
- Example

$$wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$

- This problem of less than perfect collinearity is known as **multicollinearity**
- Clearly, *work experience* and *work experience*^{2} are correlated
- If *exper* increases, *exper*^{2} will also increase

Multicollinearity: Example

- Multicollinearity is a problem in regression models where the X variables are correlated with each other
- Example

$$wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$

- This problem of less than perfect collinearity is known as **multicollinearity**
- Clearly, *work experience* and *work experience*^{2} are correlated
- If *exper* increases, *exper*^{2} will also increase

Multicollinearity: Example

- Multicollinearity is a problem in regression models where the X variables are correlated with each other
- Example

$$wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$

- This problem of less than perfect collinearity is known as **multicollinearity**
- Clearly, *work experience* and *work experience*^{2} are correlated
- If *exper* increases, *exper*^{2} will also increase

Multicollinearity: Example

- Multicollinearity is a problem in regression models where the X variables are correlated with each other
- Example

$$wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$

- This problem of less than perfect collinearity is known as **multicollinearity**
- Clearly, *work experience* and *work experience*^{2} are correlated
- If *exper* increases, *exper*^{2} will also increase

Multicollinearity: Example

- Multicollinearity is a problem in regression models where the X variables are correlated with each other
- Example
$$wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$
- This problem of less than perfect collinearity is known as **multicollinearity**
- Clearly, *work experience* and *work experience*^{2} are correlated
- If *exper* increases, *exper*^{2} will also increase

Multicollinearity: Consequences

- In theory, multicollinearity is not a serious problem - OLS estimators will still have the usual BLU properties
- In practice, it is a problem

$$wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$

- Ceteris paribus condition does not hold anymore,
- e.g. β_3 is not the marginal effect of *exper*!

Multicollinearity: Consequences

- In theory, multicollinearity is not a serious problem - OLS estimators will still have the usual BLU properties
- In practice, it is a problem

$$wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$

- Ceteris paribus condition does not hold anymore,
- e.g. β_3 is not the marginal effect of *exper*!

Multicollinearity: Consequences

- In theory, multicollinearity is not a serious problem - OLS estimators will still have the usual BLU properties
- In practice, it is a problem

$$wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$

- Ceteris paribus condition does not hold anymore,
- e.g. β_3 is not the marginal effect of *exper*!

Multicollinearity: Consequences

- In theory, multicollinearity is not a serious problem - OLS estimators will still have the usual BLU properties
- In practice, it is a problem

$$wage_i = \beta_1 + \beta_2 educ_i + \beta_3 exper_i + \beta_4 exper_i^2 + u_i$$

- Ceteris paribus condition does not hold anymore,
- e.g. β_3 is not the marginal effect of *exper*!

Multicollinearity: Consequences

- At a more technical level, multicollinearity can cause more problems
  1. **Variance**$(\hat{\beta}_2)$:

$$variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

     $r^2_{X_2X_3}$ is the squared sample correlation coefficient between X_2 and X_3
     Multicollinearity $\rightarrow$ $r^2_{X_2X_3}$ correlation coefficient high $\rightarrow$ $variance(\hat{\beta}_2)$ high $\rightarrow$ Loss of efficiency/precision estimation
  2. **t value**: A larger variance means a low t statistic and therefore statistically insignificant coefficients

$$t = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

  3. **Goodness of fit** R^2: The R^2 can end up being very high in the presence of multicollinearity
  4. **Coefficient** β: The estimates of the β can become sensitive to small changes

Multicollinearity: Consequences

- At a more technical level, multicollinearity can cause more problems
  1. **Variance**($\hat{\beta}_2$):

$$variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

     $r^2_{X_2X_3}$ is the squared sample correlation coefficient between X_2 and X_3
     Multicollinearity $\rightarrow$ $r^2_{X_2X_3}$ correlation coefficient high $\rightarrow$ $variance(\hat{\beta}_2)$ high $\rightarrow$ Loss of efficiency/precision estimation
  2. **t value**: A larger variance means a low t statistic and therefore statistically insignificant coefficients

$$t = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

  3. **Goodness of fit** R^2: The R^2 can end up being very high in the presence of multicollinearity
  4. **Coefficient** β: The estimates of the β can become sensitive to small changes

Multicollinearity: Consequences

- At a more technical level, multicollinearity can cause more problems
  1. **Variance**$(\hat{\beta}_2)$:

$$variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2 X_3}}$$

     $r^2_{X_2 X_3}$ is the squared sample correlation coefficient between X_2 and X_3
     Multicollinearity $\rightarrow$ $r^2_{X_2 X_3}$ correlation coefficient high $\rightarrow$ $variance(\hat{\beta}_2)$ high $\rightarrow$ Loss of efficiency/precision estimation
  2. **t value**: A larger variance means a low t statistic and therefore statistically insignificant coefficients

$$t = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

  3. **Goodness of fit** R^2: The R^2 can end up being very high in the presence of multicollinearity
  4. **Coefficient** β: The estimates of the β can become sensitive to small changes

Multicollinearity: Consequences

- At a more technical level, multicollinearity can cause more problems
  1. **Variance**($\hat{\beta}_2$):

     $$variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

     $r^2_{X_2X_3}$ is the squared sample correlation coefficient between X_2 and X_3
     Multicollinearity $\rightarrow$ $r^2_{X_2X_3}$ correlation coefficient high $\rightarrow$ $variance(\hat{\beta}_2)$ high $\rightarrow$ Loss of efficiency/precision estimation
  2. **t value**: A larger variance means a low t statistic and therefore statistically insignificant coefficients

     $$t = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

  3. **Goodness of fit** R^2: The R^2 can end up being very high in the presence of multicollinearity
  4. **Coefficient** β: The estimates of the β can become sensitive to small changes

Multicollinearity: Consequences

- At a more technical level, multicollinearity can cause more problems
  1. **Variance**$(\hat{\beta}_2)$:

     $$variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2 X_3}}$$

     $r^2_{X_2 X_3}$ is the squared sample correlation coefficient between X_2 and X_3
     Multicollinearity $\rightarrow$ $r^2_{X_2 X_3}$ correlation coefficient high $\rightarrow$ $variance(\hat{\beta}_2)$ high $\rightarrow$ Loss of efficiency/precision estimation
  2. **t value**: A larger variance means a low t statistic and therefore statistically insignificant coefficients

     $$t = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$$

  3. **Goodness of fit** R^2: The R^2 can end up being very high in the presence of multicollinearity
  4. **Coefficient** β: The estimates of the β can become sensitive to small changes

Multicollinearity: Consequences

```
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.6683     4.2884  -3.420 0.000677 ***
S             1.8776     0.2237   8.392 5.01e-16 ***
EXP           0.9833     0.2098   4.686 3.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


lm(formula = EARNINGS ~ S + EXP + EXPSQ, data = EAWE21)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.76580    4.57953  -3.443 0.000625 ***
S             1.86928    0.22419   8.338  7.5e-16 ***
EXP           1.42785    0.68149   2.095 0.036661 *
EXPSQ        -0.03284    0.04790  -0.686 0.493280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The correlation coefficient of *EXP* and *EXPSQ* is 0.968, which is nearly perfect collinearity
- Impact of adding *EXPSQ*:
  1. *EXP* is now only significant at the 5% significance level
  2. This is because of an increase in the Std. Err. from 0.21 to 0.68
  3. The coefficient of *EXPSQ* has the anticipated negative sign, but it is not significant.
- The loss of precision is attributable to multicollinearity,

Multicollinearity: Consequences

```
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.6683     4.2884  -3.420 0.000677 ***
S             1.8776     0.2237   8.392 5.01e-16 ***
EXP           0.9833     0.2098   4.686 3.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


lm(formula = EARNINGS ~ S + EXP + EXPSQ, data = EAWE21)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.76580    4.57953  -3.443 0.000625 ***
S             1.86928    0.22419   8.338  7.5e-16 ***
EXP           1.42785    0.68149   2.095 0.036661 *
EXPSQ        -0.03284    0.04790  -0.686 0.493280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The correlation coefficient of *EXP* and *EXPSQ* is 0.968, which is nearly perfect collinearity
- Impact of adding *EXPSQ*:
  1. *EXP* is now only significant at the 5% significance level
  2. This is because of an increase in the Std. Err. from 0.21 to 0.68
  3. The coefficient of *EXPSQ* has the anticipated negative sign, but it is not significant.
- The loss of precision is attributable to multicollinearity,

Multicollinearity: Consequences

```
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.6683     4.2884  -3.420 0.000677 ***
S             1.8776     0.2237   8.392 5.01e-16 ***
EXP           0.9833     0.2098   4.686 3.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


lm(formula = EARNINGS ~ S + EXP + EXPSQ, data = EAWE21)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.76580    4.57953  -3.443 0.000625 ***
S             1.86928    0.22419   8.338  7.5e-16 ***
EXP           1.42785    0.68149   2.095 0.036661 *
EXPSQ        -0.03284    0.04790  -0.686 0.493280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The correlation coefficient of *EXP* and *EXPSQ* is 0.968, which is nearly perfect collinearity
- Impact of adding *EXPSQ*:
  1. *EXP* is now only significant at the 5% significance level
  2. This is because of an increase in the Std. Err. from 0.21 to 0.68
  3. The coefficient of *EXPSQ* has the anticipated negative sign, but it is not significant.
- The loss of precision is attributable to multicollinearity,

Multicollinearity: Consequences

```
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.6683     4.2884  -3.420 0.000677 ***
S             1.8776     0.2237   8.392 5.01e-16 ***
EXP           0.9833     0.2098   4.686 3.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


lm(formula = EARNINGS ~ S + EXP + EXPSQ, data = EAWE21)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.76580    4.57953  -3.443 0.000625 ***
S             1.86928    0.22419   8.338  7.5e-16 ***
EXP           1.42785    0.68149   2.095 0.036661 *
EXPSQ        -0.03284    0.04790  -0.686 0.493280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The correlation coefficient of *EXP* and *EXPSQ* is 0.968, which is nearly perfect collinearity
- Impact of adding *EXPSQ*:
  1. *EXP* is now only significant at the 5% significance level
  2. This is because of an increase in the Std. Err. from 0.21 to 0.68
  3. The coefficient of *EXPSQ* has the anticipated negative sign, but it is not significant.
- The loss of precision is attributable to multicollinearity,

Multicollinearity: Consequences

```
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.6683     4.2884  -3.420 0.000677 ***
S             1.8776     0.2237   8.392 5.01e-16 ***
EXP           0.9833     0.2098   4.686 3.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


lm(formula = EARNINGS ~ S + EXP + EXPSQ, data = EAWE21)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.76580    4.57953  -3.443 0.000625 ***
S             1.86928    0.22419   8.338  7.5e-16 ***
EXP           1.42785    0.68149   2.095 0.036661 *
EXPSQ        -0.03284    0.04790  -0.686 0.493280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The correlation coefficient of *EXP* and *EXPSQ* is 0.968, which is nearly perfect collinearity
- Impact of adding *EXPSQ*:
  1. *EXP* is now only significant at the 5% significance level
  2. This is because of an increase in the Std. Err. from 0.21 to 0.68
  3. The coefficient of *EXPSQ* has the anticipated negative sign, but it is not significant.
- The loss of precision is attributable to multicollinearity,

Multicollinearity: Consequences

```
lm(formula = EARNINGS ~ S + EXP, data = EAWE21)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.6683     4.2884  -3.420 0.000677 ***
S             1.8776     0.2237   8.392 5.01e-16 ***
EXP           0.9833     0.2098   4.686 3.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


lm(formula = EARNINGS ~ S + EXP + EXPSQ, data = EAWE21)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.76580    4.57953  -3.443 0.000625 ***
S             1.86928    0.22419   8.338  7.5e-16 ***
EXP           1.42785    0.68149   2.095 0.036661 *
EXPSQ        -0.03284    0.04790  -0.686 0.493280
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The correlation coefficient of *EXP* and *EXPSQ* is 0.968, which is nearly perfect collinearity
- Impact of adding *EXPSQ*:
    1. *EXP* is now only significant at the 5% significance level
    2. This is because of an increase in the Std. Err. from 0.21 to 0.68
    3. The coefficient of *EXPSQ* has the anticipated negative sign, but it is not significant.
- The loss of precision is attributable to multicollinearity,

Multicollinearity: Detection

- How can we detect multicollinearity?
- A simple test is to calculate pairwise correlation coefficients between the X variables in the model
- High correlation coefficient values would be a first sign of the potential presence of multicollinearity

```
> df <- data.frame(wages$wage,wages$educ,wages$exper,wages$expersq)
> cor(df)
              wages.wage wages.educ wages.exper wages.expersq
wages.wage    1.00000000  0.4059033   0.1129034    0.03023781
wages.educ    0.40590333  1.0000000  -0.2995418   -0.33125594
wages.exper   0.11290344 -0.2995418   1.0000000    0.96097091
wages.expersq 0.03023781 -0.3312559   0.9609709    1.00000000
```

- However, it is not very clear what value of the pairwise correlation coefficient is considered to be too high
- Different researchers may adopt different cut-off points (e.g. $> 0.8; > 0.85; > 0.9$)
- **Limitations**: Pairwise correlation coefficients only calculate correlations between two variables.

Multicollinearity: Detection

- How can we detect multicollinearity?
- A simple test is to calculate pairwise correlation coefficients between the X variables in the model
- High correlation coefficient values would be a first sign of the potential presence of multicollinearity

```
> df <- data.frame(wages$wage,wages$educ,wages$exper,wages$expersq)
> cor(df)
              wages.wage wages.educ wages.exper wages.expersq
wages.wage    1.00000000  0.4059033   0.1129034    0.03023781
wages.educ    0.40590333  1.0000000  -0.2995418   -0.33125594
wages.exper   0.11290344 -0.2995418   1.0000000    0.96097091
wages.expersq 0.03023781 -0.3312559   0.9609709    1.00000000
```

- However, it is not very clear what value of the pairwise correlation coefficient is considered to be too high
- Different researchers may adopt different cut-off points (e.g. $> 0.8; > 0.85; > 0.9$)
- **Limitations**: Pairwise correlation coefficients only calculate correlations between two variables.

Multicollinearity: Detection

- How can we detect multicollinearity?
- A simple test is to calculate pairwise correlation coefficients between the X variables in the model
- High correlation coefficient values would be a first sign of the potential presence of multicollinearity

```
> df <- data.frame(wages$wage,wages$educ,wages$exper,wages$expersq)
> cor(df)
              wages.wage wages.educ wages.exper wages.expersq
wages.wage    1.00000000  0.4059033   0.1129034    0.03023781
wages.educ    0.40590333  1.0000000  -0.2995418   -0.33125594
wages.exper   0.11290344 -0.2995418   1.0000000    0.96097091
wages.expersq 0.03023781 -0.3312559   0.9609709    1.00000000
```

- However, it is not very clear what value of the pairwise correlation coefficient is considered to be too high
- Different researchers may adopt different cut-off points (e.g. $> 0.8; > 0.85; > 0.9$)
- **Limitations**: Pairwise correlation coefficients only calculate correlations between two variables.

Multicollinearity: Detection

- How can we detect multicollinearity?
- A simple test is to calculate pairwise correlation coefficients between the X variables in the model
- High correlation coefficient values would be a first sign of the potential presence of multicollinearity

```
> df <- data.frame(wages$wage,wages$educ,wages$exper,wages$expersq)
> cor(df)
              wages.wage wages.educ wages.exper wages.expersq
wages.wage    1.00000000  0.4059033   0.1129034    0.03023781
wages.educ    0.40590333  1.0000000  -0.2995418   -0.33125594
wages.exper   0.11290344 -0.2995418   1.0000000    0.96097091
wages.expersq 0.03023781 -0.3312559   0.9609709    1.00000000
```

- However, it is not very clear what value of the pairwise correlation coefficient is considered to be too high
- Different researchers may adopt different cut-off points (e.g. $> 0.8; > 0.85; > 0.9$)
- **Limitations**: Pairwise correlation coefficients only calculate correlations between two variables.

Multicollinearity: Detection

- How can we detect multicollinearity?
- A simple test is to calculate pairwise correlation coefficients between the X variables in the model
- High correlation coefficient values would be a first sign of the potential presence of multicollinearity

```
> df <- data.frame(wages$wage,wages$educ,wages$exper,wages$expersq)
> cor(df)
              wages.wage wages.educ wages.exper wages.expersq
wages.wage    1.00000000  0.4059033   0.1129034    0.03023781
wages.educ    0.40590333  1.0000000  -0.2995418   -0.33125594
wages.exper   0.11290344 -0.2995418   1.0000000    0.96097091
wages.expersq 0.03023781 -0.3312559   0.9609709    1.00000000
```

- However, it is not very clear what value of the pairwise correlation coefficient is considered to be too high
- Different researchers may adopt different cut-off points (e.g. $> 0.8; > 0.85; > 0.9$)
- **Limitations**: Pairwise correlation coefficients only calculate correlations between two variables.

Multicollinearity: Detection

- How can we detect multicollinearity?
- A simple test is to calculate pairwise correlation coefficients between the X variables in the model
- High correlation coefficient values would be a first sign of the potential presence of multicollinearity

```
> df <- data.frame(wages$wage,wages$educ,wages$exper,wages$expersq)
> cor(df)
              wages.wage wages.educ wages.exper wages.expersq
wages.wage    1.00000000  0.4059033   0.1129034    0.03023781
wages.educ    0.40590333  1.0000000  -0.2995418   -0.33125594
wages.exper   0.11290344 -0.2995418   1.0000000    0.96097091
wages.expersq 0.03023781 -0.3312559   0.9609709    1.00000000
```

- However, it is not very clear what value of the pairwise correlation coefficient is considered to be too high
- Different researchers may adopt different cut-off points (e.g. $> 0.8; > 0.85; > 0.9$)
- **Limitations**: Pairwise correlation coefficients only calculate correlations between two variables.

Multicollinearity: Possible Solutions

- What can you do about multicollinearity if you encounter it?
- We will discuss some possible measures, looking at the model with two explanatory variables.
- While coefficients still will be unbiased, they will have unsatisfactorily large variances.
- An obvious solution is to use methods to reduce the variances

Multicollinearity: Possible Solutions

- What can you do about multicollinearity if you encounter it?
- We will discuss some possible measures, looking at the model with two explanatory variables.
- While coefficients still will be unbiased, they will have unsatisfactorily large variances.
- An obvious solution is to use methods to reduce the variances

Multicollinearity: Possible Solutions

- What can you do about multicollinearity if you encounter it?
- We will discuss some possible measures, looking at the model with two explanatory variables.
- While coefficients still will be unbiased, they will have unsatisfactorily large variances.
- An obvious solution is to use methods to reduce the variances

Multicollinearity: Possible Solutions

- What can you do about multicollinearity if you encounter it?
- We will discuss some possible measures, looking at the model with two explanatory variables.
- While coefficients still will be unbiased, they will have unsatisfactorily large variances.
- An obvious solution is to use methods to reduce the variances

Multicollinearity: Possible Solutions

- What can you do about multicollinearity if you encounter it?
- We will discuss some possible measures, looking at the model with two explanatory variables.
- While coefficients still will be unbiased, they will have unsatisfactorily large variances.
- A obvious solution is to use methods to **reduce the variances**

Multicollinearity: Possible Solutions

- What can you do about multicollinearity if you encounter it?
- We will discuss some possible measures, looking at the model with two explanatory variables.
- While coefficients still will be unbiased, they will have unsatisfactorily large variances.
- A obvious solution is to use methods to **reduce the variances**

Multicollinearity: Possible Solutions

- What can you do about multicollinearity if you encounter it?
- We will discuss some possible measures, looking at the model with two explanatory variables.
- While coefficients still will be unbiased, they will have unsatisfactorily large variances.
- A obvious solution is to use methods to **reduce the variances**

Multicollinearity: Possible Solutions

- What can you do about multicollinearity if you encounter it?
- We will discuss some possible measures, looking at the model with two explanatory variables.
- While coefficients still will be unbiased, they will have unsatisfactorily large variances.
- A obvious solution is to use methods to **reduce the variances**

Review: Efficiency / Precision

- Multiple regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2 X_3}}$$

- $r^2_{X_2 X_3}$ is the squared sample correlation coefficient between X_2 and X_3
- Multicollinearity $\rightarrow$ $r^2_{X_2 X_3}$ correlation coefficient high $\rightarrow$ $variance(\hat{\beta}_2)$ high $\rightarrow$ Loss of efficiency/precision

Review: Efficiency / Precision

- Multiple regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2 X_3}}$$

- $r^2_{X_2 X_3}$ is the squared sample correlation coefficient between X_2 and X_3
- Multicollinearity $\rightarrow$ $r^2_{X_2 X_3}$ correlation coefficient high $\rightarrow$ $variance(\hat{\beta}_2)$ high $\rightarrow$ Loss of efficiency/precision

Review: Efficiency / Precision

- Multiple regression model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

$$variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2 X_3}}$$

- $r^2_{X_2 X_3}$ is the squared sample correlation coefficient between X_2 and X_3
- Multicollinearity $\rightarrow$ $r^2_{X_2 X_3}$ correlation coefficient high $\rightarrow$ $variance(\hat{\beta}_2)$ high $\rightarrow$ Loss of efficiency/precision

Multicollinearity: Possible Solutions

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

- Solution 1: increase n —- increase the number of observations
- Solution 2: decrease $\sigma^2_{u_i}$ —- include further relevant variables in the model
- Solution 3: increase $MSD(X_2)$
- Solution 4: decrease $r^2_{X_2X_3}$ —- combine the correlated variables

- Solution 5: Drop some of the correlated variables
- Solution 6: Theoretical restrictions

Multicollinearity: Possible Solutions

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

- Solution 1: increase n —- increase the number of observations
- Solution 2: decrease $\sigma^2_{u_i}$ —- include further relevant variables in the model
- Solution 3: increase $MSD(X_2)$
- Solution 4: decrease $r^2_{X_2X_3}$ —- combine the correlated variables

- Solution 5: Drop some of the correlated variables
- Solution 6: Theoretical restrictions

Multicollinearity: Possible Solutions

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

- Solution 1: increase n —- increase the number of observations
- Solution 2: decrease $\sigma^2_{u_i}$ —- include further relevant variables in the model
- Solution 3: increase $MSD(X_2)$
- Solution 4: decrease $r^2_{X_2X_3}$ —- combine the correlated variables

- Solution 5: Drop some of the correlated variables
- Solution 6: Theoretical restrictions

Multicollinearity: Possible Solutions

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

- Solution 1: increase n —- increase the number of observations
- Solution 2: decrease $\sigma^2_{u_i}$ —- include further relevant variables in the model
- Solution 3: increase $MSD(X_2)$
- Solution 4: decrease $r^2_{X_2X_3}$ —- combine the correlated variables
- Solution 5: Drop some of the correlated variables
- Solution 6: Theoretical restrictions

Multicollinearity: Possible Solutions

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2 X_3}}$$

- Solution 1: increase n —- increase the number of observations
- Solution 2: decrease $\sigma^2_{u_i}$ —- include further relevant variables in the model
- Solution 3: increase $MSD(X_2)$
- Solution 4: decrease $r^2_{X_2 X_3}$ —- combine the correlated variables

- Solution 5: Drop some of the correlated variables
- Solution 6: Theoretical restrictions

Multicollinearity: Possible Solutions

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

- Solution 1: increase n —- increase the number of observations
- Solution 2: decrease $\sigma^2_{u_i}$ —- include further relevant variables in the model
- Solution 3: increase $MSD(X_2)$
- Solution 4: decrease $r^2_{X_2X_3}$ —- combine the correlated variables

- Solution 5: Drop some of the correlated variables
- Solution 6: Theoretical restrictions

Multicollinearity: Possible Solutions

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

- Solution 1: increase n —- increase the number of observations
- Solution 2: decrease $\sigma^2_{u_i}$ —- include further relevant variables in the model
- Solution 3: increase $MSD(X_2)$
- Solution 4: decrease $r^2_{X_2X_3}$ —- combine the correlated variables

- Solution 5: Drop some of the correlated variables
- Solution 6: Theoretical restrictions

Multicollinearity: Possible Solution 1

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{\uparrow nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

- Solution 1: increase n —- increase the number of observations. For example,
  - Surveys: increase the budget, use clustering.
  - Time series: use quarterly instead of annual data.

```
lm(formula = S ~ ASVABC + SM + SF, data = EAWE21)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.28846    0.28347   36.29  < 2e-16 ***
ASVABC       1.23488    0.05563   22.20  < 2e-16 ***
SM           0.14780    0.02228    6.63  < 2e-11 ***
SF           0.15275    0.01971    7.75  < 7e-15 ***

nobs = 2274

lm(formula = S ~ ASVABC + SM + SF, data = EAWE21)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.59674    0.61428  17.251  < 2e-16 ***
ASVABC       1.24253    0.12359  10.054  < 2e-16 ***
SM           0.09135    0.04593   1.989   0.0473 *
SF           0.20289    0.04251   4.773  2.4e-06 ***

nobs = 500
```

Multicollinearity: Possible Solution 2

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\downarrow\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}}$$

- Solution 2: decrease $\sigma^2_{u_i}$ —- include further relevant variables in the model

```
Analysis of Variance Table

Response: S
           Df  Sum Sq Mean Sq F value    Pr(>F)
ASVABC      1 1007.00 1007.00 202.381 < 2.2e-16 ***
SM          1  112.38  112.38  22.585 2.638e-06 ***
SF          1  115.68  115.68  23.248 1.898e-06 ***
MALE        1   55.98   55.98  11.251 0.0008567 ***
Residuals 495 2462.99    4.98

Analysis of Variance Table

Response: S
           Df  Sum Sq Mean Sq F value    Pr(>F)
ASVABC      1 1007.00 1007.00 198.283 < 2.2e-16 ***
SM          1  112.38  112.38  22.128 3.312e-06 ***
SF          1  115.68  115.68  22.778 2.396e-06 ***
Residuals 496 2518.97    5.08
```

Multicollinearity: Possible Solution 3

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)\uparrow} \times \frac{1}{1 - r^2_{X_2X_3}}$$

- Solution 3: increase $MSD(X_2)$
- This is possible only at the design stage of a survey.
- Example: When planning a household survey to investigate how expenditure patterns vary with income, make sure that the sample includes a mixture of rich, poor households and middle-income households.

Multicollinearity: Possible Solution 4

- **Target: Reduce the variances**

$$\downarrow variance(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X_2)} \times \frac{1}{1 - r^2_{X_2X_3}\downarrow}$$

- Solution 4: decrease $r^2_{X_2X_3}$ —- combine the correlated variables.
- For example, create an average measure of different test scores

Multicollinearity: Possible Solution 5

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

- Solution 5: Drop some of the correlated variables
- X_2 and X_3 have higher correlation coefficient, drop X_2 or X_3
- This approach can be dangerous! Can lead to omitted variable bias
- Will be discussed in Econometrics II

Multicollinearity: Possible Solution 6

- Solution 6: Theoretical restrictions
- Think back to our educational attainment function:

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i$$

- The educational attainment will depend on the education level of the parents.
- Due to assertive matching we can assume that

$$\beta_3 = \beta_4$$

- Therefore, defining *SP* to be the sum of *SM* and *SF*, the equation may be rewritten as shown. The problem caused by the correlation between *SM* and *SF* has been eliminated

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 (SM_i + SF_i) + u_i$$

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SP_i + u_i$$

Multicollinearity: Possible Solution 6

- Solution 6: Theoretical restrictions
- Think back to our educational attainment function:

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i$$

- The educational attainment will depend on the education level of the parents.
- Due to assertive matching we can assume that

$$\beta_3 = \beta_4$$

- Therefore, defining *SP* to be the sum of *SM* and *SF*, the equation may be rewritten as shown. The problem caused by the correlation between *SM* and *SF* has been eliminated

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 (SM_i + SF_i) + u_i$$

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SP_i + u_i$$

Multicollinearity: Possible Solution 6

- Solution 6: Theoretical restrictions
- Think back to our educational attainment function:

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i$$

- The educational attainment will depend on the education level of the parents.
- Due to assertive matching we can assume that

$$\beta_3 = \beta_4$$

- Therefore, defining *SP* to be the sum of *SM* and *SF*, the equation may be rewritten as shown. The problem caused by the correlation between *SM* and *SF* has been eliminated

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 (SM_i + SF_i) + u_i$$

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SP_i + u_i$$

Multicollinearity: Possible Solution 6

- Solution 6: Theoretical restrictions
- Think back to our educational attainment function:

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i$$

- The educational attainment will depend on the education level of the parents.
- Due to assertive matching we can assume that

$$\beta_3 = \beta_4$$

- Therefore, defining *SP* to be the sum of *SM* and *SF*, the equation may be rewritten as shown. The problem caused by the correlation between *SM* and *SF* has been eliminated

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 (SM_i + SF_i) + u_i$$

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SP_i + u_i$$

Multicollinearity: Possible Solution 6

- Solution 6: Theoretical restrictions
- Think back to our educational attainment function:

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SM_i + \beta_4 SF_i + u_i$$

- The educational attainment will depend on the education level of the parents.
- Due to assertive matching we can assume that

$$\beta_3 = \beta_4$$

- Therefore, defining SP to be the sum of SM and SF, the equation may be rewritten as shown. The problem caused by the correlation between SM and SF has been eliminated

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 (SM_i + SF_i) + u_i$$

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SP_i + u_i$$

Multicollinearity: Possible Solution 6

- Solution 6: Theoretical restrictions

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 (SM_i + SF_i) + u_i$$

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SP_i + u_i$$

```
> EAWE21$SP <- EAWE21$SM +EAWE21$SF
> sfit3 <- lm(S~ASVABC+SP, data=EAWE21)
> summary(sfit3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.50285    0.61170  17.170  < 2e-16 ***
ASVABC       1.24320    0.12373  10.047  < 2e-16 ***
SP           0.15008    0.02299   6.529 1.64e-10 ***

> summary(sfit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.59674    0.61428  17.251  < 2e-16 ***
ASVABC       1.24253    0.12359  10.054  < 2e-16 ***
SM           0.09135    0.04593   1.989   0.0473 *
SF           0.20289    0.04251   4.773  2.4e-06 ***
```

- After introducing theoretical constriction, we see that the standard error is much smaller!
- Problem of multicollinearity has been eliminated.
- However, you will have to test whether this restriction is valid or not.

Multicollinearity: Possible Solution 6

- Solution 6: Theoretical restrictions

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 (SM_i + SF_i) + u_i$$

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SP_i + u_i$$

```
> EAWE21$SP <- EAWE21$SM +EAWE21$SF
> sfit3 <- lm(S~ASVABC+SP, data=EAWE21)
> summary(sfit3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.50285    0.61170  17.170  < 2e-16 ***
ASVABC       1.24320    0.12373  10.047  < 2e-16 ***
SP           0.15008    0.02299   6.529 1.64e-10 ***

> summary(sfit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.59674    0.61428  17.251  < 2e-16 ***
ASVABC       1.24253    0.12359  10.054  < 2e-16 ***
SM           0.09135    0.04593   1.989   0.0473 *
SF           0.20289    0.04251   4.773  2.4e-06 ***
```

- After introducing theoretical constriction, we see that the standard error is much smaller!
- Problem of multicollinearity has been eliminated.
- However, you will have to test whether this restriction is valid or not.

Multicollinearity: Possible Solution 6

- Solution 6: Theoretical restrictions

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 (SM_i + SF_i) + u_i$$

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SP_i + u_i$$

```
> EAWE21$SP <- EAWE21$SM +EAWE21$SF
> sfit3 <- lm(S~ASVABC+SP, data=EAWE21)
> summary(sfit3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.50285    0.61170  17.170  < 2e-16 ***
ASVABC       1.24320    0.12373  10.047  < 2e-16 ***
SP           0.15008    0.02299   6.529 1.64e-10 ***

> summary(sfit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.59674    0.61428  17.251  < 2e-16 ***
ASVABC       1.24253    0.12359  10.054  < 2e-16 ***
SM           0.09135    0.04593   1.989   0.0473 *
SF           0.20289    0.04251   4.773  2.4e-06 ***
```

- After introducing theoretical constriction, we see that the standard error is much smaller!
- Problem of multicollinearity has been eliminated.
- However, you will have to test whether this restriction is valid or not.

Multicollinearity: Possible Solution 6

- Solution 6: Theoretical restrictions

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 (SM_i + SF_i) + u_i$$

$$S_i = \beta_1 + \beta_2 ASVABC_i + \beta_3 SP_i + u_i$$

```
> EAWE21$SP <- EAWE21$SM +EAWE21$SF
> sfit3 <- lm(S~ASVABC+SP, data=EAWE21)
> summary(sfit3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.50285    0.61170  17.170  < 2e-16 ***
ASVABC       1.24320    0.12373  10.047  < 2e-16 ***
SP           0.15008    0.02299   6.529 1.64e-10 ***

> summary(sfit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.59674    0.61428  17.251  < 2e-16 ***
ASVABC       1.24253    0.12359  10.054  < 2e-16 ***
SM           0.09135    0.04593   1.989   0.0473 *
SF           0.20289    0.04251   4.773  2.4e-06 ***
```

- After introducing theoretical constriction, we see that the standard error is much smaller!
- Problem of multicollinearity has been eliminated.
- However, you will have to test whether this restriction is valid or not.