Multiple Linear Regression: Brief Overview

Multiple Linear Regression

Suppose we want to investigate the relation between:

- an outcome variable y
- a set of K explanatory variable $x_1, x_2, \dots, x_K$

The Multiple Linear Regression model assumes that the following relation is true in the population:

Population model

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_K + u_i$

- β_0 is the value of y when all x's and u equal zero
- β_k is the effect of a unit-change in x_k on y; it measures the marginal effect of x_k on y keeping u and the other x's fixed

Multiple Linear Regression

- Adding other explanatory variables improves the goodness of fit
- Makes the Zero conditional mean assumption more credible:

$$wage = \beta_0 + \beta_1 education + \beta_2 motivation + u$$

- Allows for other specifications:

$$wage = \beta_0 + \beta_1 age + \beta_2 age^2 + u$$

Assumptions

- H_1 Linearity in parameters: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u$
- H_2 Random sampling: $(y_i, x_{i1}, \ldots, x_{iK})_{i=1,\ldots,N}$ *iid*
- H_3 No perfect collinearity: $x_{ik} \neq c$ and no linear relationship between the x's in the sample
- H_4 Zero conditional mean: $E(u|x_1, \ldots, x_K) = E(u) = 0$
- H_5 Homoscedasticity: $Var(u_i|x_1, \ldots, x_K) = \sigma^2$, $\forall i = 1, \ldots, N$

Estimation

Estimation by OLS = minimizing the sum of squared residuals:

$$\sum_{i=1}^{N} \hat{u}_i^2 = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_K x_{Ki})^2$$

$K + 1$ parameters to estimate: $K + 1$ first order conditions:

$$\sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_K x_{Ki}) = 0$$

$$\sum_{i=1}^{N} x_{1i} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_K x_{Ki}) = 0$$

$$\cdots$$

$$\sum_{i=1}^{N} x_{Ki} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_K x_{Ki}) = 0$$

Unique solution under H_1 to H_4

Interpretation of OLS estimates

- $\hat{\beta}_0$,$\hat{\beta}_1, \ldots, \hat{\beta}_K$ are partial effects
- $\hat{\beta}_k$ is the effect of x_k on y after the effect of other x's on x_k is partialled out
- $\hat{\beta}_k$ is the effect of x_k on y, holding other x's fixed
- Example:

$$wage = \beta_0 + \beta_1 age + \beta_2 education + u$$

- $\hat{\beta}_2$ is the effect of schooling on wage that is not due to age, but only to the level of schooling
- Ceteris paribus = everything else equal
- $\hat{\beta}_2$ is the effect of schooling when everything else (age here) is fixed, and only the level of schooling varies

OLS in matrix form

- With multiple explanatory variables, it is easier to write the model in matrix form:

$$y = X\beta + u$$

where

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} X = \begin{pmatrix} 1 & x_{11} & \dots & x_{K1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1N} & \dots & x_{KN} \end{pmatrix} \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} u = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix}$$

OLS in matrix form

The K+1 first order conditions are simply:

$$X'(y - X\hat{\beta}) = 0$$

Which gives

$$\hat{\beta} = (X'X)^{-1}X'y$$

OLS in matrix form

Unbiasedness follows from

$$E[\hat{\beta}|X] = E[(X'X)^{-1}X'(X\beta + u)|X] = \beta + (X'X)^{-1}X'E[u|X] = \beta$$

under H_1 to H_4

From H_5, the variance of the estimator is given by:

$$\begin{aligned} V[\hat{\beta}|X] &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] \\ &= E[(X'X)^{-1}X'uu'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'E[uu'|X]X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

Estimator of σ^2

Unbiased estimator of σ^2

Under H_1 to H_5,

$$\hat{\sigma}^2 = \frac{1}{N-K-1} \sum_{i=1}^{N} \hat{u}_i^2$$

is an unbiased estimator of the error variance

So that an unbiased estimator of the variance of $\hat{\beta}$ is $\hat{\sigma}^2(X'X)^{-1}$

OLS is BLUE

Gauss-Markov Theorem

Under assumption H_1 to H_5, $\hat{\beta}$ is the best linear unbiased estimator (BLUE) of β

- B Best = smallest variance
- L Linear = $\hat{\beta}$ is a linear combination of y
- U Unbiased = $E[\hat{\beta}|X] = \beta$
- E Estimator

Multiple Regression Model: Elaboration

Multiple regression model definition and advantages

- Multiple regression model has several independent variables (also called regressors).
- Multiple regression incorporates more independent variables into the model, i.e. it is more realistic.
- It explicitly holds other factors fixed (and included in the regression) that otherwise will be in the error term.
- It allows for more flexible functional forms (including squared variables, interactions, logs, etc.)

Terminology

Multiple regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

y is dependent variable, $x_1, x_2, \ldots, x_k$ are independent variables, u is error, $\beta_0, \beta_1, .., \beta_k$ are parameters. There are k independent variables.

x_j denotes any of the independent variables, and β_j is its parameter, $j = 1 \ldots k$

Estimated equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

$\hat{y}$ is predicted value, $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_k$ are coefficients,

n is the number of observations,

i is any of the observations, $i = 1 \ldots n$.

Population	Sample
Parameter β	Coefficient $\hat{\beta}$
Error u	Residual $\hat{u}$

Residual: $\hat{u} = y - \hat{y}$

$\hat{u}$ = actual value minus predicted/fitted value for the dependent variable

Interpretation of coefficients

$$\hat{\beta}_j = \frac{\Delta y}{\Delta x_j} = \frac{change\ in\ y}{change\ in\ x_j}$$

- The coefficient $\hat{\beta}_j$ measures by how much the dependent variable changes when the independent variable x_j increases by one unit, when holding other factors fixed.
- It is assumed that the unobservable factors do not change if the independent variable changes, $\frac{\Delta u}{\Delta x_j} = 0.$

Regression model example

- Multiple regression model explaining how the wage is explained by education, experience, and tenure.
- Multiple regression model:
$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$
- Wage is measured in \$/hour, education in years, experience in years, and tenure in this company in years.
- Estimated equation for the predicted value of wage:
$$\widehat{wage} = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 exper + \hat{\beta}_3 tenure$$
- Residuals: $\hat{u} = wage - \widehat{wage}$
- We estimate the regression model to find the coefficients.
- $\hat{\beta}_1$ measures the change in wage associated with one more year of education, holding other factors fixed.

Estimated equation and interpretation

- Estimated equation

$$\begin{aligned}\widehat{wage} &= \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 exper + \hat{\beta}_3 tenure \\ &= -2.87 + 0.60\, educ + 0.02 exper + 0.17 tenure\end{aligned}$$

- Wage is measured in $/hour, education in years, experience in years, and tenure in this company in years.
- Interpretation of $\hat{\beta}_1$: the hourly wage increases by $0.60 for each additional year of education, holding other factors fixed.
- Interpretation of $\hat{\beta}_2$: the hourly wage increases by $0.02 for each additional year of experience, holding other factors fixed.
- Interpretation of $\hat{\beta}_3$: the hourly wage increases by $0.17 for each additional year of tenure in the company, holding other factors fixed.
- Interpretation of $\hat{\beta}_0$: if all regressors are zero, a person's wage is -$2.87 (but no one in the sample has zero for all regressors).

Stata output for multiple regression

```
. regress wage educ exper tenure

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(3, 522)       =     76.87
       Model |  2194.11162         3  731.370541   Prob > F        =    0.0000
    Residual |  4966.30269       522  9.51398982   R-squared       =    0.3064
-------------+----------------------------------   Adj R-squared   =    0.3024
       Total |  7160.41431       525  13.6388844   Root MSE        =    3.0845

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5989651   .0512835    11.68   0.000     .4982176    .6997126
       exper |   .0223395   .0120568     1.85   0.064    -.0013464    .0460254
      tenure |   .1692687   .0216446     7.82   0.000     .1267474    .2117899
       _cons |  -2.872735   .7289643    -3.94   0.000    -4.304799   -1.440671
------------------------------------------------------------------------------
```

The coefficients are estimated with the Stata program.

Regression model example

	Reg with 1 variable	Reg with 2 variables	Reg with 3 variables
VARIABLES	wage	wage	wage
educ	0.541***	0.644***	0.599***
	(0.0532)	(0.0538)	(0.0513)
exper		0.0701***	0.0223*
		(0.0110)	(0.0121)
tenure			0.169***
			(0.0216)
Constant	-0.905	-3.391***	-2.873***
	(0.685)	(0.767)	(0.729)

When education increases by 1 year, wage increases by $0.60, holding other factors fixed.
The coefficient on education changes slightly from one model to the next depending on what other independent variables are included in the model.

Derivation of the OLS estimates

For a regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

- We need to estimate the regression equation:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$

and find the coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_k$ by looking at the residuals

- $\hat{u} = y - \hat{y} = y - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k)$
- Obtain a random sample of data with *n* observations

(x_{ij}, y_i), where $i = 1 \ldots n$ is the observation and $j = 1 \ldots k$ is the index for the independent variable.

Derivation of the OLS estimates

- The goal is to obtain as good fit as possible of the estimated regression equation.
- Minimize the sum of squared residuals:

$$\min \sum_{i=1}^{n} \hat{u}^2 = \sum_{i=1}^{n} (y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)\,)^2$$

- We obtain the OLS coefficients $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$
- OLS is Ordinary Least Squares, based on minimizing the sum of squared residuals.

OLS properties

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 + \cdots + \hat{\beta}_k\bar{x}_k$$

The sample average of the dependent and independent variables are on the regression line.

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

The residuals sum up to zero (note that the sum of <u>squared</u> residuals are minimized).

$$\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$$

The covariance between the independent variable x_j and the residual $\hat{u}$ is zero.

Partialling out

- Partialling out shows an alternative way to obtain the regression coefficient.
- Population regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$
- First regress the independent variable on all other independent variables: $x_1 = \alpha_0 + \alpha_2 x_2 + \alpha_3 x_3 + e$
- Get the residuals from this regression: $\hat{e} = x_1 - \hat{x}_1$ (showing the part of the independent variable not explained by the other variables)
- Run the regression of the dependent variable on the residuals from this regression to obtain the same coefficient: $y = \gamma_0 + \beta_1 \widehat{e} + v$
- The coefficient $\hat{\beta}_1$ shows the relationship between y and x_1 that is not explained by the other variables. This is why “holding everything else fixed” is added when interpreting coefficients.

Partialling out example

	Dep. variable on all	Indep. variable on other	Dep. variable on
	indep. variables	indep. variables	residuals
VARIABLES	wage	educ	wage
educ	0.599***		
	(0.0513)		
exper	0.0223*	-0.0738***	
	(0.0121)	(0.00976)	
tenure	0.169***	0.0477***	
	(0.0216)	(0.0183)	
ehat			0.599***
			(0.0556)
Constant	-2.873***	13.57***	5.896***
	(0.729)	(0.184)	(0.146)

The coefficient on education is 0.599 in the original regression is the same as the coefficient on $\hat{e}$.
Interpretation: if education increases by 1 year, wage increases by $0.60, holding everything else fixed.

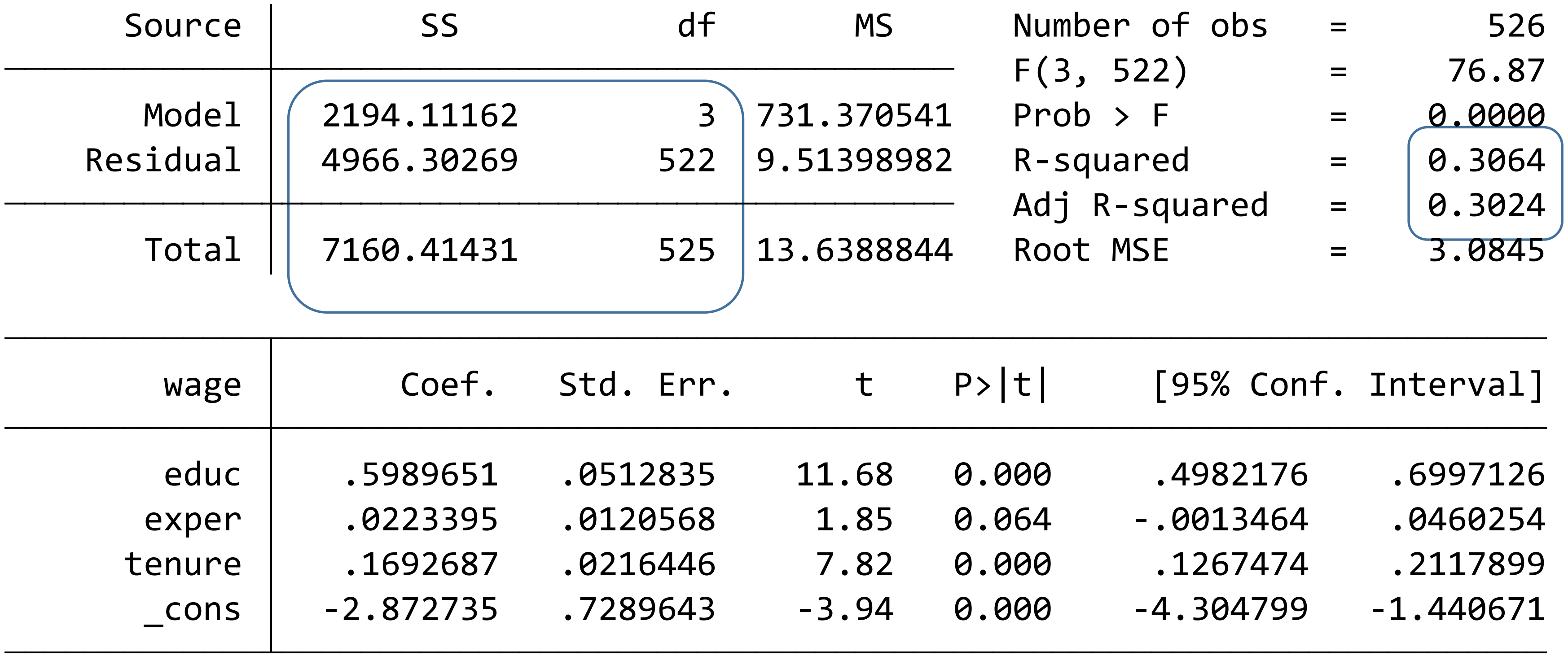
Goodness of fit measures
R-squared

- Remember that SST = SSE+ SSR or
- total sum of squares = explained sum of squares + residual sum of squares
- R-squared = R^2 = SSE/SST = 1 – SSR/SST
- R-squared is equal to the explained sum of squares divided by the total sum of squares. It measures the proportion of total variation that is explained by the regression. R-squared is a goodness of fit measure.
- An R-squared of 0.7 is interpreted as 70% of the variation is explained by the regression and the rest is due to error.
- R-squared that is greater than 0.25 is considered good fit.
- A problem with the R-squared is that it always increases when an additional regressor is added because SST is the same, but SSE increases.

Adjusted R-squared

- The adjusted R-squared is an R-squared that has been adjusted for the number of regressors in the model. The adjusted R-squared increases only if the new regressor improves the model.
- Recall that R-squared $= R^2 = 1 - \frac{SSR}{SST}$
- Adjusted R-squared = Adj $R^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$
- where *n* is the number of observations and *k* is the number of independent variables.
- Adjusted R-squared = $1 - (1 - R^2)(n - 1)/(n - k - 1)$
- As the number of regressors *k* increases, the adjusted R-squared would be penalized.
- When adding more regressors, the adjusted R^2 may increase but it also may decrease or even get negative.
- Rule: choose a model with higher adjusted R-squared when deciding whether to keep a variable or not in the model.

Adjusted R-squared calculation



```
      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(3, 522)       =     76.87
       Model |  2194.11162         3  731.370541   Prob > F        =    0.0000
    Residual |  4966.30269       522  9.51398982   R-squared       =    0.3064
-------------+----------------------------------   Adj R-squared   =    0.3024
       Total |  7160.41431       525  13.6388844   Root MSE        =    3.0845

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5989651   .0512835    11.68   0.000     .4982176    .6997126
       exper |   .0223395   .0120568     1.85   0.064    -.0013464    .0460254
      tenure |   .1692687   .0216446     7.82   0.000     .1267474    .2117899
       _cons |  -2.872735   .7289643    -3.94   0.000    -4.304799   -1.440671
------------------------------------------------------------------------------
```

R-squared = SS Model /SS Total = 2194 / 7160 = 1 – SSR/SST = 1 – 4966/7160 = 0.306

Adj R-squared = 1 – [SSR/(n-k-1)] / [SST/(n-1)] = 1 – (4966/522)/ (7160/525) = 0.302

30% of the variation in wage is explained by the regression and the rest is due to error.

R-squared and adjusted R-squared example

	Reg with 1 regressor	Reg with 2 regressor	Reg with 3 regressor
VARIABLES	wage	wage	wage
educ	0.541***	0.644***	0.599***
	(0.0532)	(0.0538)	(0.0513)
exper		0.0701***	0.0223*
		(0.0110)	(0.0121)
tenure			0.169***
			(0.0216)
Constant	-0.905	-3.391***	-2.873***
	(0.685)	(0.767)	(0.729)
SSE (explained or model)	1180	1612	2194
SST (total)	7160	7160	7160
R-squared = SSE/SST	0.165	0.225	0.306
Adj R-squared	0.163	0.222	0.302

When more regressors are added, SSE increases but SST is the same. So R-squared always increases when more variables are added (the regression can explain more of the total variation).

Adjusted R-squared can increase or decrease as more variables are added. Here, the most desirable model with the highest adj R-squared of 0.3 (explaining 30% of the variation) is the model with 3 regressors.

R-squared and adjusted R-squared examples for non-nested models

	Linear	Linear-log	Log-linear	Log-log
VARIABLES	salary	salary	lsalary	lsalary
roe	19.63*	22.67**	0.0149***	0.0179***
	(11.08)	(10.98)	(0.00433)	(0.00396)
sales	0.0163*		1.56e-05***	
	(0.00887)		(3.47e-06)	
lsales		286.3***		0.275***
		(92.33)		(0.0333)
Constant	830.6***	-1,482*	6.585***	4.362***
	(223.9)	(816.0)	(0.0875)	(0.294)
SSE (explained or model)	11,427,512	22,400,308	9	19
SST (total)	391,732,982	391,732,982	67	67
R-squared	0.029	0.057	0.129	0.282
Adjusted R-squared	0.020	0.048	0.121	0.275

SST are different for different dependent variables (salary vs lsalary). Adj R^2 for linear-log is higher than for linear model. The last model is preferred because it has the highest adjusted R-squared of 0.275.

Gauss Markov assumptions

- Gauss Markov assumptions are standard assumptions for the linear regression model

1. Linearity in parameters
2. Random sampling
3. No perfect collinearity
4. Zero conditional mean (exogeneity) – regressors are not correlated with the error term
5. Homoscedasticity – variance of error term is constant

Assumption 1: linearity in parameters

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- The relationship between the dependent and independent variables is linear in the population.
- Note that the regression model can have logged variables (e.g. log sales), squared variables (e.g. education^2) or interactions of variables (e.g. education*experience) but the β parameters are linear.

Assumption 2: random sampling

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

$(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$, where i= 1….*n*

- The data are a random sample drawn from the population. Each observation has an equal probability of being selected.
- Each observation follows the population equation.

Assumption 3: no perfect collinearity

- No perfect collinearity - none of the independent variables are constant and there are no exact relationships among them.
- An independent variable cannot be a constant because it will be collinear with the constant/intercept in the model.
- An independent variable cannot be a perfect linear combination of other independent variables (perfect collinearity). It must be dropped from the model.
- Independent variables can still be highly correlated with each other (multicollinearity), which is not a violation of this assumption though multicollinearity is also problematic.
- If independent variables are in shares that sum up to 100% (e.g. proportion of income spent on food and proportion of income not spent on food), one of the variables must be dropped. But income spent on food and income not spent on food do not sum up to 100% so both variables can be used in the model.

Perfect collinearity example

- Model with female

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + u$$

- Model with female and male

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 male + u$$

where $male = 1 - female$

This model cannot be estimated because female and male are perfectly collinear. Solution:

- Drop the collinear variable $male$
- Drop the constant (not commonly used)

Perfect collinearity example

	Model with female	Model with female and male (male is dropped)	Model with female and male but no constant
VARIABLES	wage	wage	wage
educ	0.506***	0.506***	0.506***
	(0.0504)	(0.0504)	(0.0504)
female	-2.273***	-2.273***	-1.651**
	(0.279)	(0.279)	(0.652)
o.male		-	
male			0.623
			(0.673)
Constant	0.623	0.623	
	(0.673)	(0.673)	

Multicollinearity: VIF

- Multicollinearity is when two or more independent variables are highly linearly related (correlated) with each other. The solution is to remove some of the variables or combine them.
- Multicollinearity can be checked by using variance inflation factors (VIF)
- $VIF_j = 1/(1 - R_j^2)$
- R_j^2 is the R-squared from a regression of x_j on the other independent variables.
- Higher R_j^2 would mean that x_j is well explained by other independent variables.
- From the formula above, when R_j^2 is higher, then VIF_j will be higher.
- The rule of thumb is to drop variable x_j if $VIF_j > 10$, which means that $R_j^2 > 0.9$. This variable would be well explained by the other variables in the model.
- VIF=1 for simple regression because there are no other variables and $R_j^2 = 0$
- Another rule of thumb is to drop a variable if it has a correlation coefficient above 0.9 with another variable. But it is better to use VIF for multicollinearity among all independent variables instead of correlation coefficient between two variables.

Multicollinearity example: correlation

Correlation table	Parents’ average education	Parents went to grad school	Parents went to college
Parents’ average education	1		
Parents went to grad school	0.79	1	
Parents went to college	0.81	0.42	1

Model: how test scores of a student depend on parents’ average education and whether parents went to grad school or college.

First, find the correlation between two variables at a time.

The variable of parents’ average education is highly correlated with whether parents went to grad school or college; the correlation coefficients are 0.79 and 0.81.

Multicollinearity example: VIF

VIF	Model with	Model without	Model without
	multicollinearity	multicollinearity	multicollinearity
VARIABLES	Test score	Test score	Test score
Parents’ average education	10.79		1
Parents went to grad school	4.78	1.25	
Parents went to college	4.54	1.25	

Next, estimate the models and calculate VIF.
The variable parents’ average education is highly collinear with whether parents went to grad school or college.
In the model with all variables, the VIF for parents’ average education is 10.79 which is above 10.
We need to drop this variable.
The VIFs are below 10 for models with either parents’ average education OR parents went to grad school or college, so they are OK.

Multicollinearity example

	Model with	Model without	Model without
	multicollinearity	multicollinearity	multicollinearity
VARIABLES	Test score	Test score	Test score
Parents’ average education	203.9***		148.2***
	(18.78)		(5.847)
Parents went to grad school	-1.091	5.829***	
	(0.758)	(0.475)	
Parents went to college	-2.426***	2.648***	
	(0.587)	(0.350)	
Constant	162.7***	545.1***	251.4***

The coefficients in the models without and with multicollinearity are very different.
In the model with multicollinearity we can wrongly conclude that if parents went to college, the test scores of the student will be lower.
Drop the variable that is causing multicollinearity (parents’ average education) and use last two models.
Standard errors are higher for model with multicollinearity.

Assumption 4: zero conditional mean (exogeneity)

$$E(u_i|x_{i1}, x_{i2}, \ldots, x_{ik}) = 0$$

- Expected value of error term u given the independent variables x is zero.
- The expected value of the error must not differ based on the values of the independent variables.
- With more variables, this assumption is more likely to hold because fewer unobservable factors are in the error term.
- Independent variables that are correlated with the error term are called endogenous; endogeneity is a violation of this assumption.
- Independent variables that are uncorrelated with the error term are called exogenous; this assumption holds if all independent variables are exogenous.

Zero conditional mean example

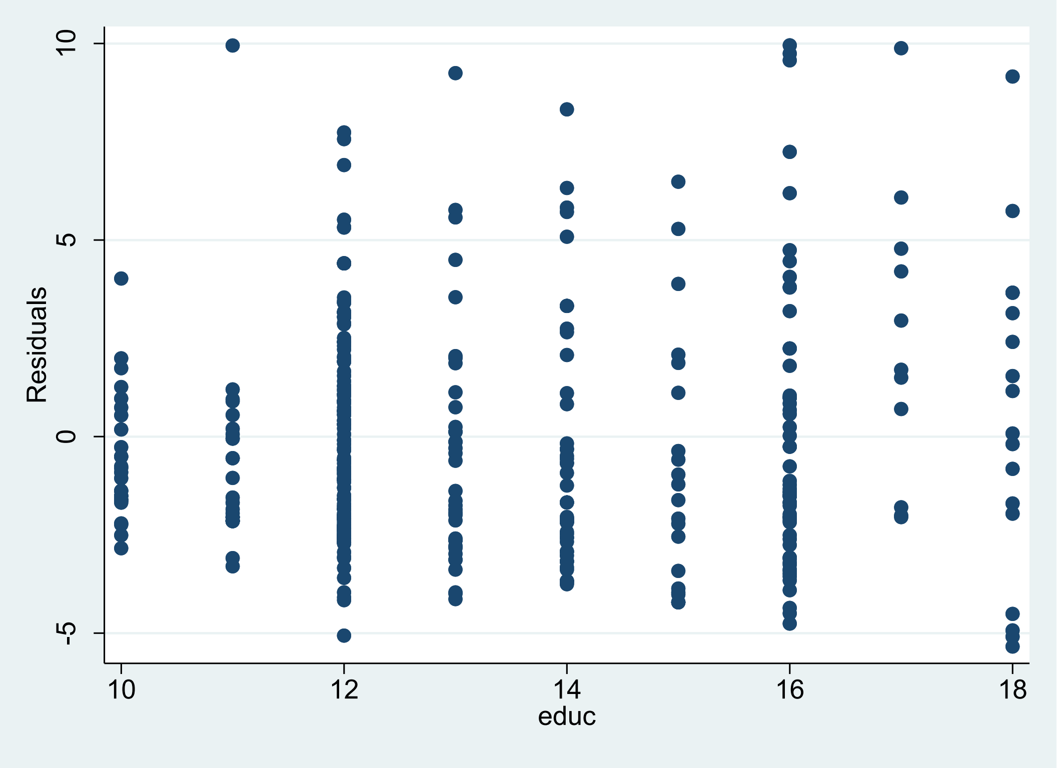
Regression model

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + u$$

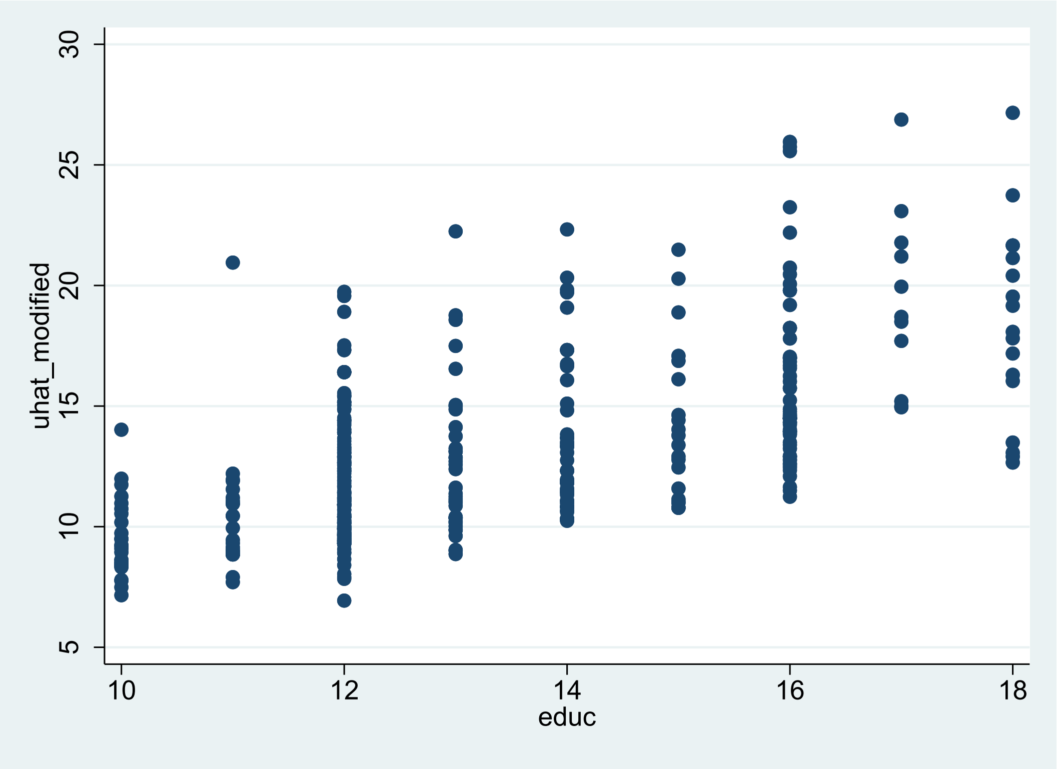
- In the example of wage and education, when ability (which is unobserved and part of the error) is higher, education would also be higher.
- This is a violation of the zero conditional mean assumption.

Example of exogeneity vs endogeneity

Exogeneity - zero conditional mean



Endogeneity - conditional mean is not zero



E(u|x)=0 error term is the same given education

E(u|x)>0 ability/error is higher when education is higher

Unbiasedness of the OLS estimators

- Gauss Markov Assumptions 1-4 (linearity, random sampling, no perfect collinearity, and zero conditional mean) lead to the unbiasedness of the OLS estimators.

  $E(\hat{\beta}_j) = \beta_j$, where $j = 0, 1, \ldots, k$
- Expected values of the sample coefficients $\hat{\beta}$ are the population parameters β.
- If we estimate the regression model with many random samples, the average of these coefficients will be the population parameter.
- For a given sample, the coefficients may be very different from the population parameters.

Omitted variable bias

- The "true" population regression model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
- We need to estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
- But instead we estimate a misspecified model: $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$, where x_2 is the omitted variable from this model.
- If $x_1$and x_2 are correlated, there will be a relationship between them

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

Substitute in above equation to get:

$$y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u$$
$$= (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) x_1 + (\beta_2 v + u\ )$$

The coefficient that will be estimated for x_1 when x_2 is omitted will be biased.

Omitted variable bias

- An unbiased coefficient is when $E\left(\hat{\beta}_1\right) = \beta_1$, but this coefficient is biased because $E\left(\tilde{\beta}_1\right) = \beta_1 + \beta_2\ \delta_1$, where $\beta_2\delta_1$ is the bias.
- With an omitted variable, the coefficient will not be biased if
  - $\beta_2 = 0$. Looking at $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, this means that x_2 does not belong in the model (x_2 is irrelevant).
  - $\delta_1 = 0$. Looking at $x_2 = \delta_0 + \delta_1 x_1 + v$, this means that x_2 and $x_1$are not correlated.
  - In other words, if the omitted variable is irrelevant $\beta_2 = 0$ or uncorrelated $\delta_1 = 0$, there will be no omitted variable bias.

Omitted variable bias example

- Suppose the “true” population regression model is:

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u$$

- But instead we estimate a misspecified model:

$$wage = \alpha_0 + \alpha_1 educ + e$$

where $abil$ is the omitted variable from this model.

- If $abil$ and $educ$ are correlated, there will be a relationship between them

$$abil = \delta_0 + \delta_1 educ + v$$

Substitute in above equation to get:

$$\begin{aligned} wage &= \beta_0 + \beta_1 educ + \beta_2(\delta_0 + \delta_1 educ + v) + u \\ &= (\beta_0+\beta_2\delta_0) + (\beta_1+\beta_2\delta_1)educ + (\beta_2 v + u\ ) \end{aligned}$$

When $abil$ is omitted, the coefficient that will be estimated for $educ$ is $\tilde{\beta}_1$= $\beta_1 + \beta_2\delta_1$, where $\beta_2\delta_1$ is the bias.

Omitted variable bias example

	Model with educ and abil	Model with educ (abil is omitted)	Model of abil and educ
VARIABLES	wage	wage	abil
educ	1.153***	1.392***	0.551***
	(0.127)	(0.103)	(0.0213)
abil	0.433***		
	(0.137)		
Constant	-2.523	-4.857***	-5.389***
	(1.543)	(1.360)	(0.282)

Coefficient in model with omitted variable = $\tilde{\beta}_1 \mathrm{biased}$ =
= original coefficient + bias = $\hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}_1$ = 1.153 + 0.433*0.551 = 1.153 + 0.238 = 1.392
In the model with omitted variable, for each additional year of education, wage increases by $1.39 instead of $1.15, so there is a bias of $0.24.
The effect of education on wage is overestimated because people with higher education have higher ability.
The coefficient on education is also reflecting the $0.24 effect of ability on wage through its relationship with edu

Omitted variable bias example

	Model with educ and exper	Model with educ (exper is omitted)	Model of exper and educ
VARIABLES	wage	wage	exper
educ	1.948***	1.392***	-0.905***
	(0.139)	(0.103)	(0.0275)
exper	0.614***		
	(0.105)		
Constant	-18.70***	-4.857***	22.54***
	(2.724)	(1.360)	(0.364)

Now, suppose that we have experience instead of ability for the true model.

Coefficient in model with omitted variable = 1.392 = original coefficient + bias = 1.948 + 0.614*(-0.905)= 1.948-0.556

In the model with omitted variable, for each additional year of education, wage increases by $1.39 instead of $1.95, so there is a negative bias of -$0.56.

The effect of education on wage is underestimated because people with higher education have less experience.

The coefficient on education is also reflecting the -$0.56 effect of experience on wage through its relationship with educ.

Including irrelevant variables

- Suppose the "true" population regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- But instead we estimate a misspecified model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

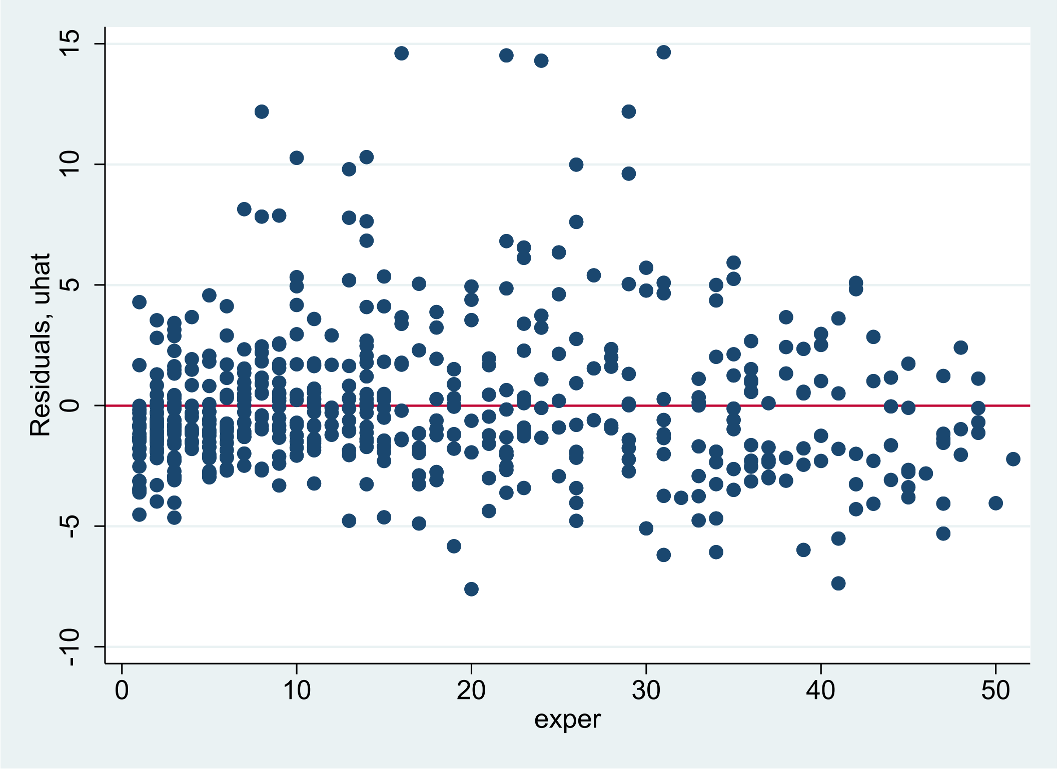
where x_3 is an irrelevant variable that is included.

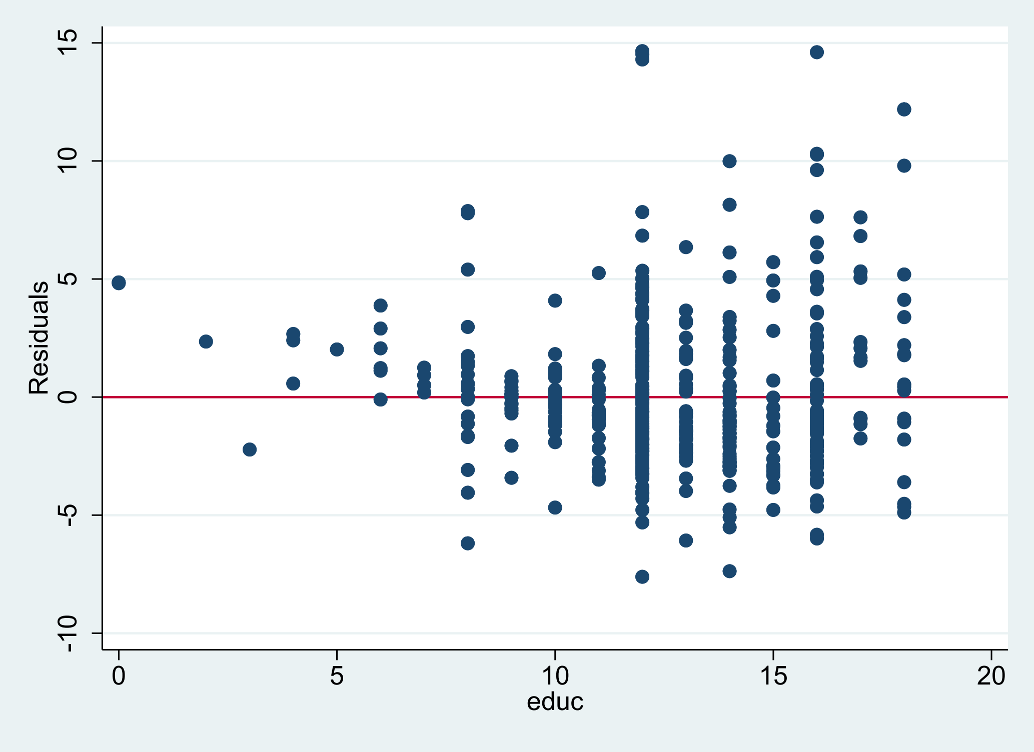
- In the population regression model, $E\left(\hat{\beta}_3\right) = \beta_3 = 0$ because the variable is not included.
- Adding an irrelevant variable does not cause bias but it increases the sampling variance.

Assumption 5: homoscedasticity

- Homoscedasticity $var(u_i|x_{i1}, x_{i2}, \dots, x_{ik}) = \sigma^2$
- Variance of the error term u must not differ with the independent variables.
- Heteroscedasticity $var(u_i|x_{i1}, x_{i2}, \dots, x_{ik}) \neq \sigma^2$is when the variance of the error term u is not constant for each x.

Homoscedasticity vs heteroscedasticity example





Homoscedasticity for experience

$var(u|exper) = \sigma^2$

Heteroscedasticity for education

$var(u|educ) \neq \sigma^2$

Variances of the OLS estimators

Recall with simple regression: $var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{SST_x}$

With multiple regression: $var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{SST_j(1-R_j^2)}$

- $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sampling variance of variable x_j. High variance of the regressor will lead to lower variance of the coefficient.
- R_j^2 is the R-squared from a regression of x_j on all other independent variables. High R-squared means that the regressor is very well explained by other regressors (multicollinearity) and it will have higher variance of its coefficient.
- σ^2 is the variance of the error term. High variance in the error term will lead to high variance of the coefficients.
- Estimators with lower variance are also called more precise and are desirable. This means that high variance in the independent variable, low R-squared for x_j (less multicollinearity), and low variance in the error term are desirable.

Unbiasedness of the error variance

- The variance of the error term σ^2 is not known but can be estimated as the variance of the residuals, corrected for the number of regressors k.

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} \hat{u}_i^2$$

- The degrees of freedom is $(n-k-1)$.
- Gauss Markov Assumptions 1-5 (linearity, random sampling, no perfect collinearity, zero conditional mean, and homoscedasticity) lead to the unbiasedness of the error variance.

$$E(\hat{\sigma}^2) = \sigma^2$$

Standard errors of the regression coefficients

- Standard errors of the regression coefficients:

$$se\left(\hat{\beta}_j\right) = \sqrt{var\left(\hat{\beta}_j\right)} = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1 - R_j^2)}}$$

- The standard errors are the square root of the variances.
- The unknown population variance of error term σ^2 is replaced with the sample variance of the residuals $\hat{\sigma}^2$.
- The standard errors measure how precisely the coefficients are calculated. They are calculated and included in the regression output.

Variance in misspecified models

- "True" population regression model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
- We need to estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
- But instead we estimate a misspecified model: $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$, where x_2 is the omitted variable from this model.

$$var(\hat{\beta}_1) = \frac{\sigma^2}{SST_1(1-R_1^2)} > var(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1}$$

- There is a trade off between bias and variance.
- Omitted variables lead to a misspecified model, and biased coefficients, which is undesirable. But the coefficients have lower variance, which is desirable.
- Typically it is better to have unbiased coefficients first and then lower variance (precision), so it is better to include the independent variable in the model.
- But irrelevant regressors should not be included. They will not cause bias but will increase variance.

Variance in misspecified model example (same as omitted variable bias example)

	Model with educ and abil	Model with educ (abil is omitted)
VARIABLES	wage	wage
educ	1.153***	1.392***
	(0.127)	(0.103)
abil	0.433***	
	(0.137)	
Constant	-2.523	-4.857***
	(1.543)	(1.360)

The standard error of the coefficient on education is 0.103 in the misspecified model which is lower than in the original model 0.127.

Remember that the standard error is the square root of the variance.

Gauss-Markov theorem

- Gauss Markov theorem says that under assumptions 1-5 (linearity in parameters, random sampling, no perfect collinearity, zero conditional mean, homoscedasticity), the OLS estimators are best linear unbiased estimators (BLUE) of the coefficients.
- Estimators are linear in the dependent variable.

$\hat{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$, for example in simple regression $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

- From all linear estimators, OLS estimators have the lowest variance (best means lowest variance).