Applied econometrics note

Multiple Linear Regression Model

Table of Contents

Preface

This set of notes originates from my Applied Econometrics I module at UNNC. Rather than adding new material, I have reorganized the logical flow of the key points presented in the lecture slides. The module is closely linked to ECON1049 Mathematical Economics and Econometrics (SPR)—now renamed ECON1057 Introductory Econometrics—so many topics will already be familiar. If you find something related to statistics or econometrics, please go back to the lecture notes in ECON1049 (/ECON 1057). This will benefit to your memory and smooth your further study.

That said, the course should not be taken lightly. Its true value lies not in memorizing the slide-deck formulas, but in learning how to apply the statistical and econometric theory covered in Year 2 to real economic-data analysis. From a practical standpoint, I strongly recommend gaining proficiency in STATA, as intensive coding will return in Applied Econometrics II and even your final year dissertation.

Because I studied this module on exchange at UNUK, I also completed a piece of coursework. That assignment—requiring the direct application of PPT concepts, extensive STATA work, and a basic research framework—proved immensely beneficial. If you know classmates heading to UNUK, do not hesitate to ask for a look at their coursework. At UNNC, Applied Econometrics I is assessed solely by a final exam; therefore, if you are curious about econometrics or research methods, consider obtaining the UNUK coursework and experimenting with it yourself (perhaps over the summer, or even with GPT's guidance). Treat it as a scaffold for practicing real applications, but ultimately try to write the code on your own.

My guiding principle is to look beyond earning marks (though I know marks are important) for isolated knowledge points and instead ask whether each step you take builds the capabilities you will need next year or even in the future. Much like the Permanent Income Hypothesis in macroeconomics, cultivating a habit of extension and forward thinking will smooth your progression to the next stage, or if you learn repeated game in game theory, you will find out that the choices in each stage will be different depending on whether the game is one-stage or multi-stage (I think that sometimes, in real life, acting in player 1 or 2 in game theory is important, because they always have rational strategy).

This is the synthesized note 2 (4 in total). If you do not have the previous notes, please visit the website: https://github.com/wang95483/notebook. And since this is just a synthesized lecture notes, its function is limited. So if you have any other ideas of how to learn this course, feel free to drop me an email first hmyhw8@nottingham.edu.cn (or if this doesn't work, use wang95483@gmail.com , my personal email).

Three important implications of moving from simple to multiple regression analysis

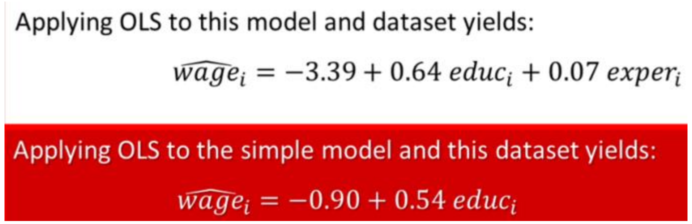
(1) The estimators change: In general, the estimators from simple & multiple regressions are **not equal.**

Formula

SLR: $$\hat{\beta}_1 = \frac{Cov(x_1,y)}{Var(x_1)}$$

MLR: $$\hat{\beta}_1 = \frac{Cov(x_1y)Var(x_2) - Cov(x_2y)Cov(x_1x_2)}{Var(x_1)Var(x_2) - Cov(x_1x_2)^2}$$

Example



Applying OLS to this model and dataset yields:

$$\widehat{wage}_i = -3.39 + 0.64\ educ_i + 0.07\ exper_i$$

Applying OLS to the simple model and this dataset yields:

$$\widehat{wage}_i = -0.90 + 0.54\ educ_i$$

Special case: equal estimator in simple linear regression model and multivariable linear regression model

- x_1 and x_2 are uncorrelated, i.e., $Cov(x_1x_2) = 0$, then the SLR and MLR estimators are the same
- the *ceteris paribus* effect of a change in x_2 on $\hat{y}$ is zero, i.e., $\hat{\beta}_2 = 0$

These results also hold for the estimators of β_2

(2) The estimators (and the estimates they generate) have a "partialling out" interpretation.

In the MLR we "partial out" that part of the variation in y that could be explained by either $x1$ or $x2$ or a mixture of both. This makes the estimation more demanding but, is an important step towards a ceteris paribus interpretation.

(3) A ceteris paribus (causal) interpretation of the estimators (and the estimates they generate) is more likely to be correct.

In the MLR, we have controlled for the effect of x_2 on y. In the SLR we have not.

→ a **ceteris paribus interpretation** of the estimator of β_1 from the MLR is more likely to be correct.

A model with two regressors

Formula

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

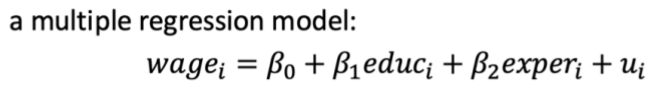
Zero conditional assumptions

For our estimators to be **unbiased**, the **zero conditional mean assumption** $E[u|x_1, x_2] = 0$, must be valid.

= for any values of x_1 and x_2 in the population, the <u>average value of u is zero</u>.

Examples

Explicit Model



a multiple regression model:

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + u_i$$

<u>Explanation</u>

experience has been **taken out of the error term** and included directly in the model

investigate the relationship between:

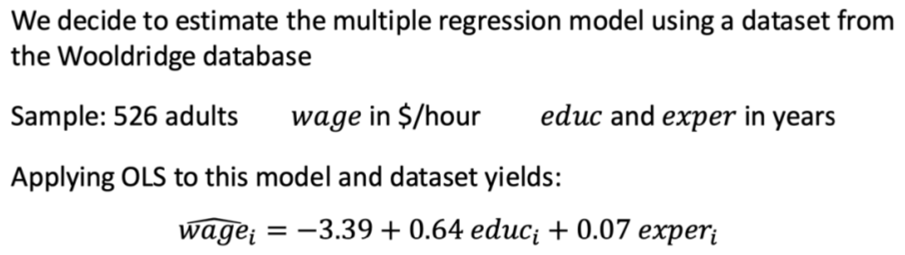
(1) experience and wages

(2) education and wages, while controlling for the effect of experience on wages

→ We can use the estimated model to predict **the effect of any combination of changes in the explanatory variables** (regressors) on the dependent variable.

(3) The zero conditional mean assumption is E[u|educ, exper] = 0

→ Add dataset



We decide to estimate the multiple regression model using a dataset from the Wooldridge database

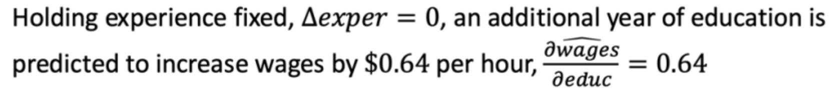
Sample: 526 adults $wage$ in \$/hour $educ$ and $exper$ in years

Applying OLS to this model and dataset yields:

$$\widehat{wage}_i = -3.39 + 0.64\ educ_i + 0.07\ exper_i$$

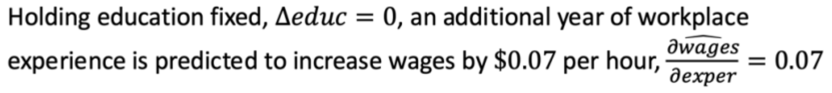
Explanation

(1) Effect of education on wages



Holding experience fixed, $\Delta exper = 0$, an additional year of education is predicted to increase wages by \$0.64 per hour, $\frac{\partial \widehat{wages}}{\partial educ} = 0.64$

(2) Effect of experience on wages



Holding education fixed, $\Delta educ = 0$, an additional year of workplace experience is predicted to increase wages by \$0.07 per hour, $\frac{\partial \widehat{wages}}{\partial exper} = 0.07$

(3) But spending an additional year of education reduces potential years of workplace experience by one.

So, what is the effect of staying in education an extra year ($\Delta educ$ = 1) and, thereby, forfeiting a year of workplace experience ($\Delta exper$ = −1)?

$$\Delta \widehat{wage}_i = 0.64\ \Delta educ_i + 0.07\ \Delta exper_i$$

$$\Delta \widehat{wage}_i = \ (0.64 \times 1)\ + (0.07 \times -1) = 0.57$$

(4) Zero conditional mean assumption: Other factors omitted from the model are uncorrelated with education and experience.

A model with k regressors

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$

Explanation

u is the **disturbance term** that represents all the factors **other than $x_1, x_2, x_3, \ldots x_k$ that affect y**

←→No matter how many regressors we have in the model there will always be **unobserved factors** that we cannot control for.

These are **collectively represented by u.**

Zero conditional assumptions

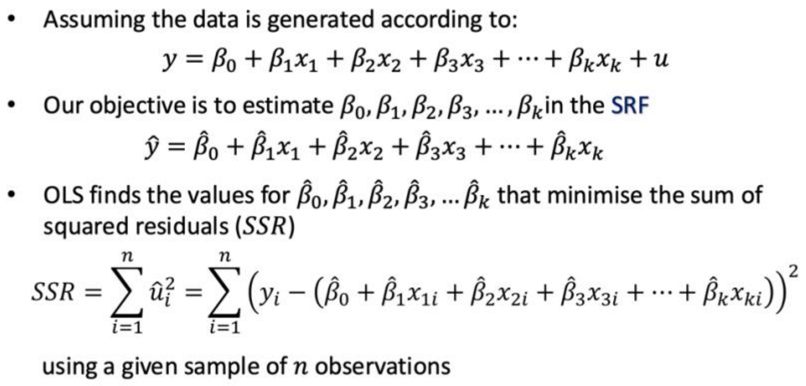
$E(u|x_1, x_2, x_3, \ldots x_k) = 0$

←→all **unobserved factors (u)** are **uncorrelated with each** and every one of the **explanatory variables**.

If any one of the explanatory variables is **correlated with u**, then no matter how large k is, the OLS estimators will be **biased.**

The mechanics of OLS

Derive OLS estimators for multivariable linear models



- Assuming the data is generated according to:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$

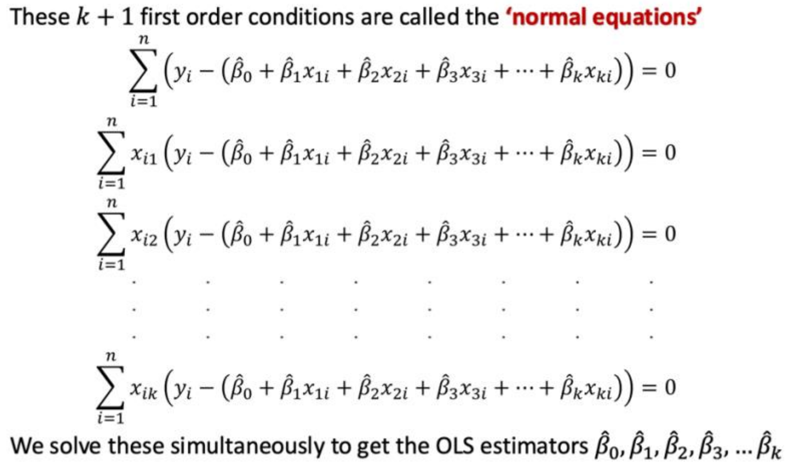
- Our objective is to estimate $\beta_0, \beta_1, \beta_2, \beta_3, \ldots, \beta_k$in the SRF

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \cdots + \hat{\beta}_k x_k$$

- OLS finds the values for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots \hat{\beta}_k$ that minimise the sum of squared residuals (SSR)

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_k x_{ki}) \right)^2$$

using a given sample of n observations



These $k+1$ first order conditions are called the **'normal equations'**

$$\sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_k x_{ki}) \right) = 0$$

$$\sum_{i=1}^{n} x_{i1} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_k x_{ki}) \right) = 0$$

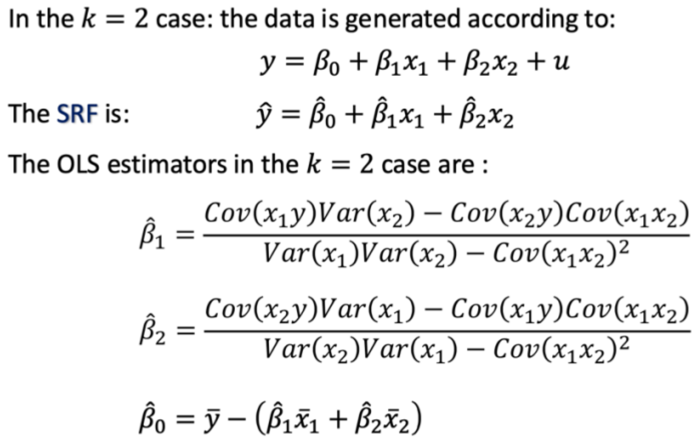
$$\sum_{i=1}^{n} x_{i2} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_k x_{ki}) \right) = 0$$

$$\vdots$$

$$\sum_{i=1}^{n} x_{ik} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_k x_{ki}) \right) = 0$$

We solve these simultaneously to get the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots \hat{\beta}_k$

Example



In the $k = 2$ case: the data is generated according to:

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

The SRF is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

The OLS estimators in the $k = 2$ case are :

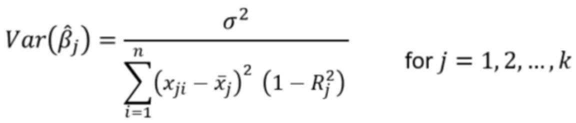
$$\hat{\beta}_1 = \frac{Cov(x_1 y)Var(x_2) - Cov(x_2 y)Cov(x_1 x_2)}{Var(x_1)Var(x_2) - Cov(x_1 x_2)^2}$$

$$\hat{\beta}_2 = \frac{Cov(x_2 y)Var(x_1) - Cov(x_1 y)Cov(x_1 x_2)}{Var(x_2)Var(x_1) - Cov(x_1 x_2)^2}$$

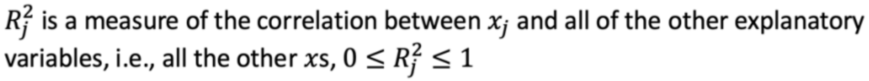
$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2)$$

Variance of the OLS estimators in the MLR (after Gauss-Markov assumptions)

The variances of the OLS estimators



$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2 \; (1 - R_j^2)} \qquad \text{for } j = 1, 2, \ldots, k$$



R_j^2 is a measure of the correlation between x_j and all of the other explanatory variables, i.e., all the other xs, $0 \le R_j^2 \le 1$

The variances are:

(1) larger the larger σ^2, the variance of the error (disturbance) terms
(2) smaller the larger n, the size of the sample used in the estimation
(3) smaller the greater the variance in the explanatory variables $\boldsymbol{x}$
(4) larger the greater the correlation between the explanatory variables $\boldsymbol{x}$. (the problem of multicollinearity)

- The greater correlation between x_1 and x_2 the closer the R_1^2 gets to its maximum value of 1
- The closer R_1^2 gets to 1, the closer $Var(\hat{\beta}_1)$ gets to ∞
- If $R_1^2 = 1$, x_1 and x_2 are perfectly collinear (MLR3 violated)

Estimating $Var(\hat{\beta}_j)$

$$\begin{aligned}\hat{u}_i &= y_i - \hat{y}_i \\ &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki})\end{aligned}$$

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

$\hat{\sigma}^2$ is an unbiased estimator of σ^2

→

$$Var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2 \, (1 - R_j^2)} \qquad \text{for } j = 1, 2, \ldots, k$$

→

(1) The standard error of the regression

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

(2) The standard errors of the OLS estimators

$$se(\hat{\beta}_j) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (x_{ji} - \bar{x}_j)^2 \, (1 - R_j^2)}}$$

Multicollinearity

It **does not** cause **bias** in the estimators.

It only seriously **inflates** $Var(\hat{\beta}_j)$ when the correlation very **high.**

NOTE: in STATA, we use $se(\hat{\beta})$ to illustrate this phenomenon.

It is tempting to remove one (or more) of the explanatory variables, but this will lead to bias in the estimators.

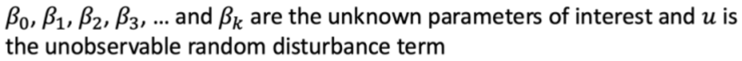
Solution: **Increasing** the **sample size**, n.

The Gauss-Markov Assumptions

Assumption MLR1: linear in parameters

The (true) population model is a linear function of the explanatory variables.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u$$



β_0, β_1, β_2, β_3, ... and β_k are the unknown parameters of interest and u is the unobservable random disturbance term

A regression model is said to be linear in parameters if the dependent variable is expressed as a **linear combination of the parameters (coefficients).**

i.e. $\beta_0, \beta_1, \ldots, \beta_k$ are not raised to any power, multiplied together, or transformed.

Assumption MLR2: random sampling

The sample of size n, $\{(y_i, x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ik})$: i = 1, 2, ... , $n\}$, taken from the population and used to generate the estimates, is **random** and **$n > k + 1$.**

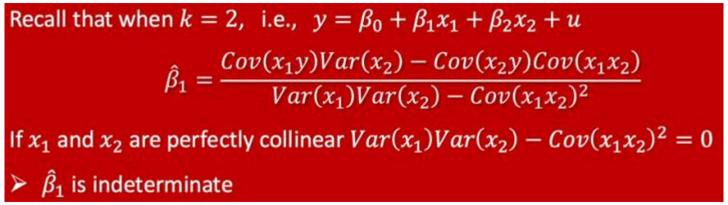
Assumption MLR3: no perfect collinearity (形容的是 variable 之间的，不是 variable 和 y 之间的)

(1) There is **variation** in all the variables.
(2) How the explanatory variables relate to one another:

There are **no exact linear relationships** among the **independent variables** (reflects the fact that there is more than one explanatory variable).

←→it makes no sense to measure the effect of changes in a variable that does not change, there is no point in including a variable that is a perfect linear combination of another.

If there the assumption is invalid, then **estimation is impossible.**



Recall that when $k = 2$, i.e., $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

$$\hat{\beta}_1 = \frac{Cov(x_1 y)Var(x_2) - Cov(x_2 y)Cov(x_1 x_2)}{Var(x_1)Var(x_2) - Cov(x_1 x_2)^2}$$

If x_1 and x_2 are perfectly collinear $Var(x_1)Var(x_2) - Cov(x_1 x_2)^2 = 0$

➢ $\hat{\beta}_1$ is indeterminate

Solution: exclude one of the explanatory variables

Difference between multicollinearity and perfect collinearity

	multicollinearity	**perfect collinearity**
Definition	High correlation between variables: Multicollinearity occurs when two or more independent variables in a regression model are highly correlated but not perfectly correlated. This means one variable can be approximately explained by a linear combination of other variables.	**Exact linear relationship** between variables
OLS coefficients	Estimable but imprecise: Large standard errors for the coefficients. Difficulty in interpreting individual effects of correlated variables.	Perfect linearity causes the OLS estimation to fail. So it cannot be uniquely estimated.
Phenomenon	inflates standard errors	

Assumption MLR4: zero conditional mean

How the explanatory variables relate to the disturbance term, $\boldsymbol{u}$:

The disturbance term, u, has an expected value of zero given any values of the independent variables.

←→ $E\ (u|x_1,\, x_2,\, x_3,\, \ldots\ ,\, x_k)$ = 0

If this assumption is violated, the estimation is possible, but the estimators will be **biased.**

Zero conditional mean assumption is more likely to be valid in the multiple regression model because more variables are explicitly included in the model.

Reason

When an independent variable is correlated with u, it means that part of the variation in X is due to omitted variables, measurement errors, or other issues captured in the error term. OLS attributes some of this variation to X, causing the coefficient to be incorrect.

Bias in Coefficient Estimates

If an independent variable (X) is correlated with the error term (u), the OLS estimator for its coefficient (β) will be biased.

This means the estimated coefficient will not accurately reflect the true effect of X on the dependent variable (Y).

The bias occurs because OLS cannot distinguish the variation in Y caused by X from the variation caused by the error term.

The greater the correlation between X and u, the larger the bias in $\hat{\beta}$

Unreliable Interpretation

The coefficient of X will likely overestimate or underestimate the true relationship, depending on the direction of the correlation with u.

This makes it impossible to trust the causal interpretation of the coefficient.

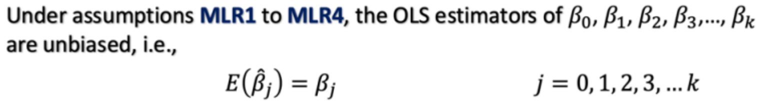
Standard Errors

The standard errors of the coefficients may also be distorted, leading to incorrect inference (e.g., misleading p-values or confidence intervals).

"The omitted variable bias is **downwards**" means that the estimated coefficient for a variable in a regression model is *smaller (closer to zero) than the true coefficient* due to the omission of a relevant variable.

MLR1-MLR4 ➔

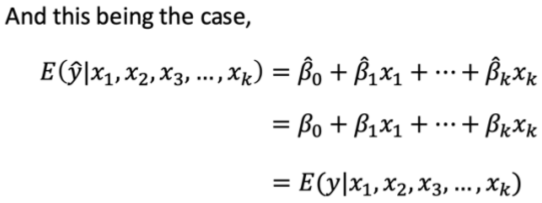
(1)



Under assumptions **MLR1** to **MLR4**, the OLS estimators of $\beta_0, \beta_1, \beta_2, \beta_3, \ldots, \beta_k$ are unbiased, i.e.,

$$E\left(\hat{\beta}_j\right) = \beta_j \qquad\qquad j = 0, 1, 2, 3, \ldots k$$

(2)



And this being the case,

$$\begin{aligned} E(\hat{y}|x_1, x_2, x_3, \ldots, x_k) &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \\ &= E(y|x_1, x_2, x_3, \ldots, x_k) \end{aligned}$$

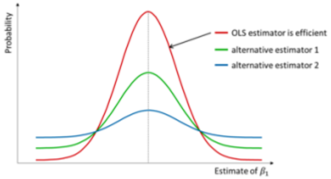
Assumption MLR5: homoscedasticity

The errors (disturbances) have the same (constant) variance **irrespective** of the values of the explanatory variables.

$$Var(u|x_1, x_2, x_3, \dots, x_k) = \sigma^2$$
$$Var(u|\boldsymbol{x}) = \sigma^2$$

MLR1-MLR5 →

Best	(smallest variance)
Linear	(linear function of the data)
Unbiased	$(E(\hat{\beta}_j) = \beta_j)$
Estimators	


Probability
OLS estimator is efficient
alternative estimator 1
alternative estimator 2
Estimate of β_1