Introductory Econometrics

Statistical Properties of the OLS Estimator, Interpretation of OLS Estimates and Effects of Rescaling

Monash Econometrics and Business Statistics

Semester 2, 2018

Recap

- The multiple regression model can be written in matrix form as

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times(k+1)}{\mathbf{X}} \underset{(k+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{u}}$$

- The OLS procedure finds a linear combination of $\mathbf{X}$ that is closest to the vector $\mathbf{y}$, i.e. the length of its error vector (OLS residuals) is the shortest
- This implies that the OLS residual vector is perpendicular to all columns of $\mathbf{X}$, i.e.,

$$\mathbf{X}'\hat{\mathbf{u}} = 0$$

  which leads to the famous OLS formula

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- A consequence of orthogonality of residuals and columns of $\mathbf{X}$ is that

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{u}_i^2$$

or

$$\text{SST} = \text{SSE} + \text{SSR}$$

- This leads to the definition of the coefficient of determination R^2, which is a measure of goodness of fit

$$R^2 = \text{SSE}/\text{SST} = 1 - \text{SSR}/\text{SST}$$

- OLS formula gives us fitted values that are closest to the actual values in the sense that the length of the residual vector is the smallest possible.
- But why should this be a good estimate of the unknown parameters of the conditional expectation function?
- We ask if this formula produces the best information we can get from our data about the unknown parameters β

Lecture Outline

- Interpretation of the OLS estimates in multiple regression (textbook reference 3-2a to 3-2e)
- Properties of good estimators
- Properties of the OLS estimator $\widehat{\boldsymbol{\beta}}$
  1. OLS is unbiased: The expected value of $\widehat{\boldsymbol{\beta}}$ (Textbook reference 3-3, and Appendix E-2)
  2. The Variance of $\widehat{\boldsymbol{\beta}}$ (Textbook reference 3-4, and Appendix E-2)
  3. OLS is BLUE (Gauss-Markov Theorem): The Efficiency of $\widehat{\boldsymbol{\beta}}$ (Textbook reference 3-5, and Appendix E-2)
- Units of measurement: do the results qualitatively change if we change the units of measurement? (textbook reference 2-4a, 6-1 (exclude 6-1a))

Interpretation of OLS estimates

Example: The causal effect of education on wage

- Consider again the *wage* equation written as:

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

  where *IQ* is IQ score (in the population, it has a mean of 100 and sd = 15).

- Primarily interested in β_1, because we want to know the value that education adds to a person's wage.
- Without *IQ* in the equation, the coefficient of *educ* will show how strongly *wage* and *educ* are correlated, but both could be caused by a person's ability.
- By explicitly including *IQ* in the equation, we obtain a more persuasive estimate of the causal effect of education provided that *IQ* is a good proxy for intelligence.

Interpretation of OLS estimates

Example: The causal effect of education on wage

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	146.9524	77.71496	1.890916	0.0589
EDUC	60.21428	5.694982	10.57322	0.0000

R-squared	0.107000	Mean dependent var	957.9455
Adjusted R-squared	0.106043	S.D. dependent var	404.3608
S.E. of regression	382.3203	Akaike info criterion	14.73253
Sum squared resid	1.36E+08	Schwarz criterion	14.74289

- If we only estimate a regression of wage on education, we cannot be sure if we are measuring the effect of education, or if education is acting as a proxy for smartness. This is important, because if the education system does not add any value other than separating smart people from not so smart, the society can achieve that much cheaper by national IQ tests!

Interpretation of OLS estimates

Example: The causal effect of education on wage

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-128.8899	92.18232	-1.398206	0.1624
EDUC	42.05762	6.549836	6.421171	0.0000
IQ	5.137958	0.955827	5.375403	0.0000

R-squared	0.133853	Mean dependent var	957.9455
Adjusted R-squared	0.131995	S.D. dependent var	404.3608
S.E. of regression	376.7300	Akaike info criterion	14.70414
Sum squared resid	1.32E+08	Schwarz criterion	14.71967

- The coefficient of *educ* now shows that for two people with the same *IQ* score, the one with 1 more year of education is expected to earn $42 more.

Interpretation of OLS estimates

- Consider $k = 2$ for simplicity
- The conditional expectation of y given x_1 and x_2 (also know as the population regression function) is

$$E(y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- The estimated regression (also known as the sample regression function) is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

- The formula

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2$$

allows us to compute how predicted y changes when x_1 and x_2 change by any amount.

- What if we increase x_1 by 1 unit and hold x_2 fixed, that is, $\Delta x_1 = 1$ and $\Delta x_2 = 0$?

$$\Delta \hat{y} = \hat{\beta}_1 \text{ if } \Delta x_1 = 1 \text{ and } \Delta x_2 = 0$$

- In other words, $\hat{\beta}_1$ is the change in predicted y when x_1 increases by 1 unit while x_2 is held fixed.
- We also refer to $\hat{\beta}_1$ as the estimate of the partial effect of x_1 on y holding x_2 constant
- Yet another legitimate interpretation is that $\hat{\beta}_1$ estimates the effect of x_1 on y after the influence of x_2 has been removed (or has been controlled for)

- Similarly,
$$\Delta\hat{y} = \hat{\beta}_2 \text{ if } \Delta x_1 = 0 \text{ and } \Delta x_2 = 1$$
- Let's go back to regression output and interpret the parameters.
$$\begin{aligned} \widehat{wage} &= -128.89 + 42.06\ educ + 5.14\ IQ \\ n &= 935,\ R^2 = .134 \end{aligned}$$
- 42.06 shows that for two people with the same IQ, the one with one more year of education is predicted to earn \$42.06 more in monthly wages.
- Or: Every extra year of education increases the predicted wage by \$42.06, keeping IQ constant (or "after controlling for IQ", or "after removing the effect of IQ", or "all else constant", or "all else equal", or "*ceteris paribus*")

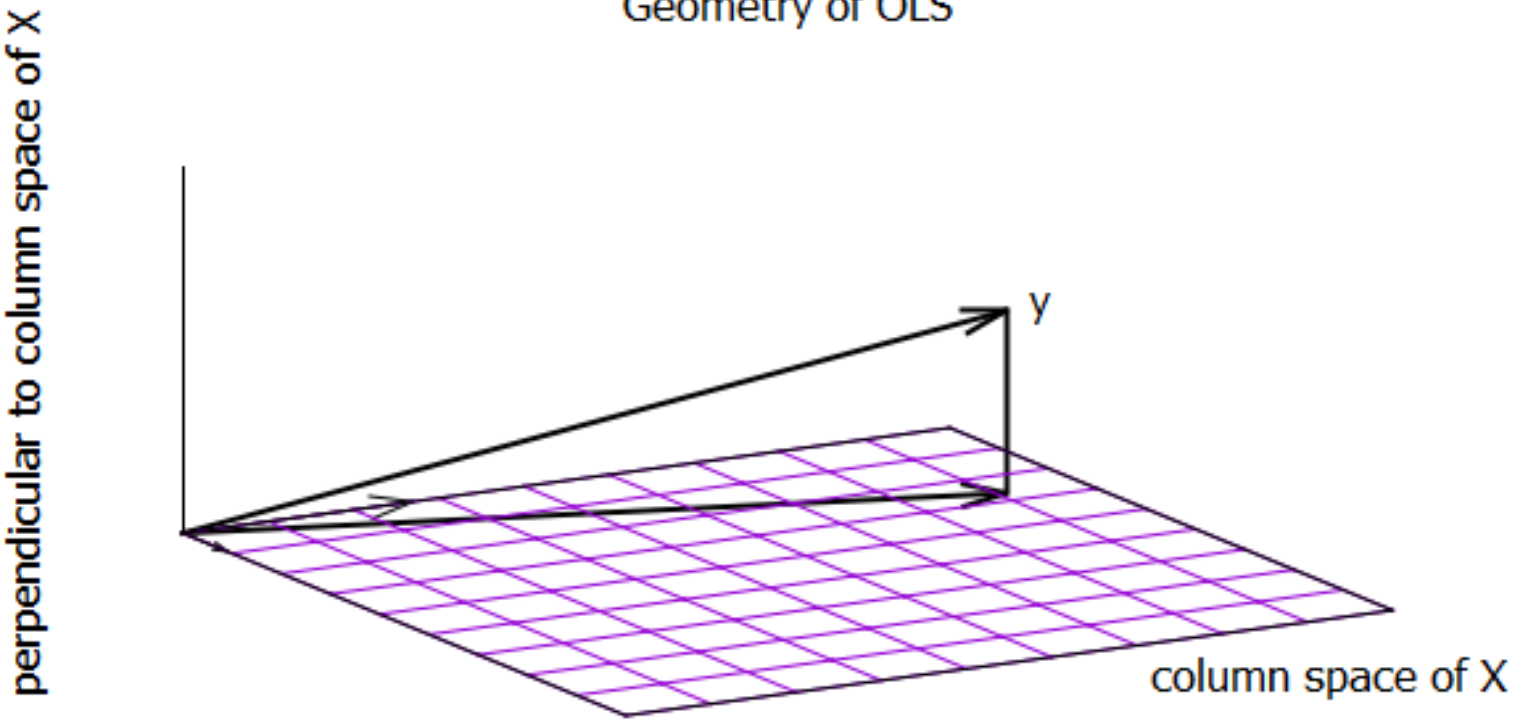
Effects of rescaling

The upshot: changing units of measurement does not create any extra information, so it only changes OLS results in predictable and non-substantive ways.

- Scaling $x \to cx$: Any change of unit of measurement that involves multiplying x by a constant (such as changing dollars to cents or pounds to kilograms), does not change the column space of $\mathbf{X}$, so will not change $\hat{\mathbf{y}}$. Therefore, it must be that the coefficient of x changes such that $x\hat{\beta}$ stays the same.

$$\begin{aligned} \hat{y} &= \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 \\ \hat{y} &= \widehat{\beta_0^*} + \widehat{\beta_1^*} x_1^* + \widehat{\beta_2^*} x_2 \\ x_1^* = c x_1 \quad &\Rightarrow \quad \widehat{\beta_0^*} = \widehat{\beta_0},\ \widehat{\beta_1^*} = \widehat{\beta_1}/c,\ \widehat{\beta_2^*} = \widehat{\beta_2} \end{aligned}$$

Geometry of OLS



- Change of units of the form $x \to a + cx$: Any change of unit of measurement that involves multiplying x by a constant and adding or subtracting a constant (such as changing from $^\circ C$ to $^\circ F$) does not change the column space of $\mathbf{X}$ either, so will not change $\hat{\mathbf{y}}$. Therefore, it must be that the $\hat{\boldsymbol{\beta}}$ changes in such a way that $\mathbf{X}\hat{\boldsymbol{\beta}}$ stays the same.

$$\begin{aligned} \hat{y} &= \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 \\ \hat{y} &= \widehat{\beta_0^*} + \widehat{\beta_1^*} x_1^* + \widehat{\beta_2^*} x_2 \\ x_1^* = a + cx_1 \quad &\Rightarrow \quad \widehat{\beta_0^*} = \widehat{\beta_0} - a\widehat{\beta_1}/c,\ \widehat{\beta_1^*} = \widehat{\beta_1}/c,\ \widehat{\beta_2^*} = \widehat{\beta_2} \end{aligned}$$

- In both cases, since $\hat{\mathbf{y}}$ does not change, residuals, SST, SSE and SSR all stay the same, so R^2 will not change.

- Changing the units of dependent variable $y \rightarrow cy$: This changes the length of $\mathbf{y}$ but the column space of $\mathbf{X}$ is still the same. So, $\hat{y}$ will be multiplied by c, and since both $\mathbf{y}$ and $\hat{\mathbf{y}}$ are multiplied by c, the residuals will be multiplied by c also.

$$\begin{aligned} y &= \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \hat{u} \\ y^* &= \widehat{\beta_0^*} + \widehat{\beta_1^*} x_1 + \widehat{\beta_2^*} x_2 + \hat{u^*} \\ y^* = cy &\Rightarrow \widehat{\beta_0^*} = c\widehat{\beta}_0,\ \widehat{\beta_1^*} = c\widehat{\beta}_1,\ \widehat{\beta_2^*} = c\widehat{\beta}_2,\ \hat{u^*} = c\hat{u} \end{aligned}$$

- In this case, SST, SSE and SSR all change but all will be multiplied by ..., so again, R^2 will not change.

Estimators and their Unbiasedness

- An estimator is a formula that combines sample information and produces an estimate for parameters of interest.
- Example: Sample average is an estimator for the population mean.
- Example: $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is an estimator for the parameter vector $\boldsymbol{\beta}$ in the multiple regression model.
- Since estimators are functions of sample observations, they are ...
- While we do not know the values of population parameters, we can use the power of mathematics to investigate if the expected value of the estimator is indeed the parameter of interest
- Definition: An estimator is an unbiased estimator of a parameter of interest if its expected value is the parameter of interest.

The Expected Value of the OLS Estimator

- Under the following assumptions, $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$

Multiple Regression Model

MLR.1 Linear in Parameters $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$	**E.1 Linear in Parameters** $\underset{n\times 1}{\mathbf{y}} = \underset{n\times(k+1)}{\mathbf{X}} \underset{(k+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{u}}$
MLR.2 Random Sampling We have a random sample of n observations	**E.3 Zero Conditional Mean**
MLR.4 Zero Conditional Mean $E(u \mid x_1, x_2, \ldots, x_k) = 0$	$E(\mathbf{u} \mid \mathbf{X}) = \underset{(n\times 1)}{\mathbf{0}}$
MLR.3 No Perfect Collinearity None of $x_1, x_2, \ldots, x_k$ is a constant and there are no *exact linear* relationships among them	**E.2 No Perfect Collinearity** $\mathbf{X}$ has rank $k+1$

- We use an important property of conditional expectations to prove that the OLS estimator is unbiased: if $E(z \mid w) = 0$, then $E(g(w)z) = 0$ for any function g. For example $E(wz) = 0$, $E(w^2 z) = 0$, etc.
- Now let's show $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
- The assumption of no perfect collinearity (E.2) is immediately required because ...
- Step 1: Using the population model (Assumption E.1), substitute for $\mathbf{y}$ in the estimator's formula and simplify

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \end{aligned}$$

- Step 2: Take expectations:

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}) &= E(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}) \\ &= \boldsymbol{\beta} + E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}) \\ &= \boldsymbol{\beta} \end{aligned}$$

  using Assumption E.3, since $E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0} \Rightarrow E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}) = \mathbf{0}$

- Note that all assumptions were needed and were used in this proof.

- Are these assumptions too strong?
- Linearity in parameters is not too strong, and does not exclude non-linear relationships between y and x (more on this later)
- Random sample is not too strong for cross-sectional data if participation in the sample is not voluntary. Randomness is obviously not correct for time series data.
- Perfect multicollinearity is quite unlikely unless we have done something silly like using income in dollars and income in cents in the list of independent variables, or we have fallen into the "dummy variable trap" (more on this later)
- Zero conditional mean is not a problem for predictive analytics, because for best prediction, we always want our best estimate of $E(y \mid x_1, x_2, \ldots, x_k)$ for a set of observed predictors.

- Zero conditional mean can be a problem for prescriptive analytics (causal analysis) when we want to establish the causal effect of one of the x variables, say x_1 on y controlling for an attribute that we cannot measure. E.g. Causal effect of education on wage keeping ability constant:

  We want to estimate β_1 in: $wage = \beta_0 + \beta_1 educ + \beta_2 ability + u$

  We run the regression: $wage = \hat{\alpha}_0 + \hat{\alpha}_1 educ + \hat{v}$

$$E(\hat{\alpha}_1) = \beta_1 + \beta_2 \frac{\partial ability}{\partial educ} \neq \beta_1$$

- $\hat{\alpha}_1$ is a biased estimator of β_1
- This is referred to as "omitted variable bias"
- One solution is to add a measurable proxy for ability, such as IQ
- Other solutions in ETC3410

- Zero conditional mean can be quite restrictive in time series analysis as well.
- It implies that the error term in any time period t is uncorrelated with each of the regressors, in *all time periods*, past, present and future.
- Assumption **MLR.4** is violated when the regression model contains a lag of the dependent variable as a regressor. E.g. We want to predict this quarter's GDP using GDP outcomes for the past 4 quarters
- In this case, the regression parameters are biased estimators.

The Variance of the OLS Estimator

- Coming up with unbiased estimators is not too hard. But how the estimates that produce are dispersed around the mean determines how precise they are.
- To study the variance of $\widehat{\boldsymbol{\beta}}$, we need to learn about variances of a vector of random variables var-cov matrix

- We introduce an extra assumption:

MLR.1 Linear in Parameters $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$	**E.1 Linear in Parameters** $\underset{n\times 1}{\mathbf{y}} = \underset{n\times(k+1)}{\mathbf{X}} \underset{(k+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{u}}$
MLR.2 Random Sampling We have a random sample of n observations	**E.3 Zero Conditional Mean** $E(\mathbf{u} \mid \mathbf{X}) = \underset{(n\times 1)}{\mathbf{0}}$
MLR.4 Zero Conditional Mean $E(u \mid x_1, x_2, \ldots, x_k) = 0$	
MLR.3 No Perfect Collinearity None of $x_1, x_2, \ldots, x_k$ is a constant and there are no *exact linear* relationships among them	**E.2 No Perfect Collinearity** $\mathbf{X}$ has rank $k+1$
MLR.5 Homoskedasticity $E(u^2 \mid x_1, x_2, \ldots, x_k) = \sigma^2$	**E.4 Homo+Randomness** $Var(\mathbf{u} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$

- Under these assumptions, the variance-covariance matrix of the OLS estimator conditional on $\mathbf{X}$ is

$$Var(\widehat{\beta} \mid \mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

because

- We can immediately see that given $\mathbf{X}$ the OLS estimator will be more precise (i.e. its variance will be smaller) the smaller σ^2 is
- It can also be seen (not as obvious) that as we add observations to the sample, the variance of the estimator decreases, which also makes sense
- Gauss-Markov Theorem: Under Assumptions E.1 to E.4 (or MLR.1 to MLR.5) $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (B.L.U.E.) of $\boldsymbol{\beta}$.
- This means that there is no other estimator that can be written as a linear combination of elements of $\mathbf{y}$ that will be unbiased and will have a lower variance than $\widehat{\boldsymbol{\beta}}$.
- This is the reason that everybody loves the OLS estimator.

Estimating the Error Variance

- We can calculate $\widehat{\beta}$, and we showed that

$$Var(\widehat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

  but we cannot compute this because we do not know σ^2

- An unbiased estimator of σ^2 is

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

- The square root of $\hat{\sigma}^2$, $\hat{\sigma}$, is reported by eviews in regression output under the name **standard error of the regression**, and it is a measure of how good the fit is (the smaller, the better)

- Why do we divide SSR by $n-k-1$ instead of n?
- In order to get an unbiased estimator of σ^2

$$E(\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k-1}) = \sigma^2.$$

  If we divide by n the expected value of the estimator will be slightly different from the true parameter (proof of unbiasedness of $\hat{\sigma}^2$ is not required)
- Of course if the sample size n is large, this bias is very small
- Think about dimensions: $\hat{\mathbf{y}}$ is in the column space of $\mathbf{X}$, so it is in a subspace with dimension $k+1$
- $\hat{\mathbf{u}}$ is orthogonal to column space of $\mathbf{X}$, so it is in a subspace with dimension $n-(k+1) = n-k-1$. So even though there are n coordinates in $\hat{\mathbf{u}}$, only $n-k-1$ of those are free (it has $n-k-1$ **degrees of freedom**)

Standard error of each parameter estimate

- The **standard error** of each $\hat{\beta}_j, j = 0, 1, \ldots, k$ is the square root of the diagonal elements of

$$\widehat{Var(\hat{\boldsymbol{\beta}} \mid \mathbf{X})} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

- These are reported under the heading **Std. Error** in eviews
- When reporting regression results, we usually report these standard errors under their corresponding parameter estimates

- Example: Predicting wage with education and experience

Dependent Variable: WAGE
Method: Least Squares
Sample: 1 526
Included observations: 526

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.390540	0.766566	-4.423023	0.0000
EDUC	0.644272	0.053806	11.97397	0.0000
EXPER	0.070095	0.010978	6.385291	0.0000

R-squared	0.225162	Mean dependent var	5.896103
Adjusted R-squared	0.222199	S.D. dependent var	3.693086
S.E. of regression	3.257044	Akaike info criterion	5.205204
Sum squared resid	5548.160	Schwarz criterion	5.229531
Log likelihood	-1365.969	Hannan-Quinn criter.	5.214729
F-statistic	75.98998	Durbin-Watson stat	1.820274
Prob(F-statistic)	0.000000		

- We report these results as

$$\widehat{wage} = \underset{(0.77)}{-3.39} + \underset{(0.05)}{0.64}\ education + \underset{(0.01)}{0.07}\ experience$$

Summary

- Under the assumptions that:
  1. the population model is linear in parameters;
  2. the conditional expectation of true errors given all explanatory variables is zero;
  3. the sample is randomly selected; and
  4. the explanatory variables are not perfectly collinear

  the OLS estimator $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.
- If we add no conditional heteroskedasticity to the above assumptions, then OLS is the B.L.U.E. for $\boldsymbol{\beta}$

Summary

- **Interpretation of OLS estimates in multiple regression:** Very important to interpret the coefficients using the context, and very important to remember that each $\widehat{\beta}$ estimates the partial effect of its corresponding x all other x staying constant.
- **Linear scaling:** Linear transformation of dependent or independent variables changes the outcomes of linear regression in exactly predictable ways, and all of these outcomes are substantively the same. Hence, we should use the units of measurement which make presentation of results more understandable without worrying that changing the units of data may affect our results.

Digression: The Variance-Covariance Matrix

- Consider n random variables $v_1, v_2, \ldots, v_n$ with means $\mu_1, \mu_2, \ldots, \mu_n$. We can write:

$$\underset{n\times 1}{\mathbf{v}} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \quad E(\mathbf{v}) = \underset{n\times 1}{\boldsymbol{\mu}} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

- Denote the variance of v_i by σ_i^2 for $i = 1, \ldots, n$, and covariance of v_i and v_j by σ_{ij} for $i = 1, \ldots, n$, $j = 1, \ldots, n$ and $i \neq j$

- We define the **variance-covariance matrix** of $\mathbf{v}$ as

$$\begin{aligned} Var(\mathbf{v}) &= E(\mathbf{v}-\boldsymbol{\mu})(\mathbf{v}-\boldsymbol{\mu})' \\ &= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} \end{aligned}$$

- Note that since $\sigma_{ij} = Cov(v_i, v_j) = Cov(v_j, v_i) = \sigma_{ji}$, this matrix is symmetric.
- An important property: If $\mathbf{A}$ is an $n \times k$ matrix of constants, then

$$Var(\mathbf{A}'\mathbf{v}) = \mathbf{A}' Var(\mathbf{v}) \mathbf{A}$$

- Always check that dimensions are compatible when using matrices

variance