Introductory Econometrics

Simple and Multiple Regression Analysis

Monash Econometrics and Business Statistics

Semester 2, 2018

Recap

- We reviewed the concepts of:
  1. a random variable and its probability density function (pdf);
  2. the mean and the variance of a random variable;
  3. the joint pdf of two random variables;
  4. the covariance and correlation between two random variables;
  5. the mean and the variance of a linear combination of random variables;
  6. how and why diversification reduces risk, and how good estimators in econometrics reduce risk in a similar fashion;
  7. conditional distribution of one random variable given another;
  8. the conditional expectation function as the fundamental object for modelling.

Lecture Outline: Wooldridge: Ch 2 - 2.3, Appendix B-1, B-2, B-4

- Explaining y using x with a model
- Simple linear regression (textbook reference Chapter 2, 2-1)
- The OLS estimator (textbook reference, Ch 2, 2-2,2-3)
- Simple linear regression in matrix form
- Geometric interpretation of least squares (not in the textbook)
- Stretching our imagination to multiple regression - Appendix E of the textbook, section E-1

What is a model?

- To study the relationship between random variables, we use a *model*
- A complete model for a random variable is a model of its probability distribution
- For example, a possible model for heights of adult men is that it is normally distributed, and since normal distribution is fully described by its mean and variance, we only have to estimate the mean and variance from a sample of observations on adult men and we have estimated a model example1
- But when studying two or more random variables, modelling the joint distribution is difficult and requires a lot of data.
- So, we model the conditional distribution of y given x.
- In particular we provide a model for $E(y \mid x)$

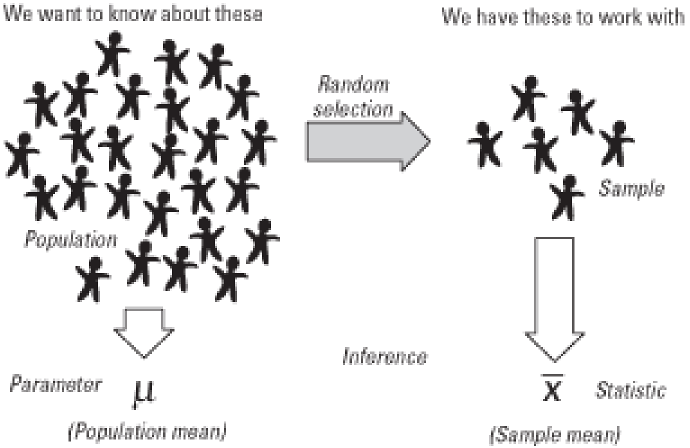
Terminology and notation

- Notation: Unlike first year, we use lower case letters for random variables
- Terminology: There are many different ways people refer to the target variable y that we want to explain using variable x. These include:

y	x
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Regressand	Regressor

- The expressions
  "Run a regression of y on x", and,
  "Regress y on x"
  both mean "Estimate the model $y = \beta_0 + \beta_1 x + u$ using the ordinary least squares method"

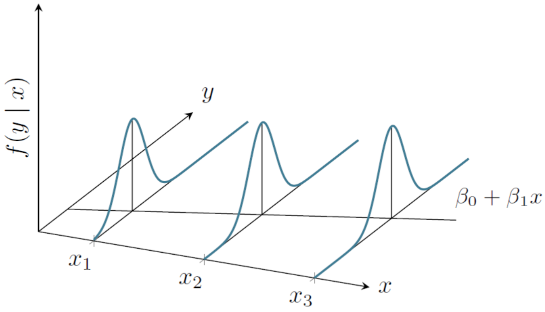
Modelling mean



$$E(y \mid x) = \beta_0 + \beta_1 x$$
(Conditional expectation function)

$$\hat{y} = \widehat{E(y \mid x)} = \hat{\beta}_0 + \hat{\beta}_1 x$$
(Sample regression function)

The simple linear regression model in a picture



The simple linear regression model in equation form

- The following equation specifies the conditional mean of y given x

$$y = \beta_0 + \beta_1 x + u \text{ with } E(u \mid x) = 0$$

- It is an incomplete model because it does not specify the probability distribution of y conditional of x
- If we add the assumptions that $Var(u \mid x) = \sigma^2$ and that the conditional distribution of u given x is normal, then we have a complete model, which is called the "classical linear model" and was shown in the picture on the previous slide.
- In summary: we make the assumption that in the big scheme of things, data are generated by this model, and we want to use observed data to learn the unknowns β_0, β_1 and σ^2 in order to predict y using x.

The OLS estimator

- Let's leave the theory universe and go back to the data world
- We have a random sample of n observations on two variables x and y in two columns of a spreadsheet
- Let's denote them by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- We want to see if these two columns of numbers are related to each other
- In the simple case that there is only one x, we look at the scatter plot, which is very informative visual tools and give us a good idea of the correlation between y and x (bodyfat and weight)
- Unfortunately though, in some business and economic applications the signal is too weak to be detected by data visualisation alone (asset return and size)

The OLS estimator

- How can we determine a straight line that fits our data best?
- We find $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimise the **sum of squared residuals**

$$SSR(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

- My favourite visual explanation of this is in http://youtu.be/jEEJNz0RK4Q
- The first order conditions are:

$$\frac{\partial SSR}{\partial b_0}\Big|_{\widehat{\beta}_0, \widehat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$
$$\frac{\partial SSR}{\partial b_1}\Big|_{\widehat{\beta}_0, \widehat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

- Two equations and two unknowns, we can solve to get $\widehat{\beta}_0$ and $\widehat{\beta}_1$

- We obtain the set of equations

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$
$$\sum_{i=1}^{n} x_i \hat{u}_i = \sum_{i=1}^{n} x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

- And after some algebra (see page 28 of the textbook, 5th ed., or page 26, 6th ed.)

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

- Several important things to note:

    1. $$\widehat{\beta}_1 = \frac{\widehat{Cov(x,y)}}{\widehat{Var(x)}} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}$$

       Those of you who are doing finance, now realise where the name "*beta*" of a stock has come from!

    2. From the formula for $\widehat{\beta}_0$ we see that $\bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}$, which shows that $(\bar{x}, \bar{y})$ lies on the regression line, i.e. regression prediction is most accurate for the sample average.

    3. Both $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are functions of sample observations only. They are estimators, i.e. formulae that can be computed from the sample. If we collect a different sample from the same population, we get different estimates. So these estimators are random variables.

Estimators we have seen so far

Population parameter	its estimator
population mean $\mu_y = E(y)$	sample average $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
population variance $\sigma_y^2 = E(y-\mu_y)^2$	sample variance $s_y^2 = \hat{\sigma}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$
population st. dev. $\sigma_y = \sqrt{\sigma_y^2}$	sample st. dev. $\hat{\sigma}_y = \sqrt{\hat{\sigma}_y^2}$
population covariance $\sigma_{xy} = E(x-\mu_x)(y-\mu_y)$	sample covariance $\hat{\sigma}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$
population correlation coeff. $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$	sample corr. coeff. $\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x\hat{\sigma}_y}$
slope and intercept of the PRF β_1 and β_0 in $E(y \mid x) = \beta_0 + \beta_1 x$	their OLS estimators $\widehat{\beta}_1 = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}$ $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x}$

Simple linear regression in matrix form

- For each of the n observations (x_1, y_1),(x_2, y_2),…,(x_n, y_n), our model implies

$$y_i = \beta_0 + \beta_1 x_i + u_i \ , \ \ i = 1, \ldots, n$$

- We stack all of these on top of each other

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \beta_0 + \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \beta_1 + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

- We define the $n \times 1$ vectors $\mathbf{y}$ (dependent variable vector) and $\mathbf{u}$ (error vector):

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \text{ and } \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

and the $n \times 2$ matrix of regressors

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

and the 2×1 parameter vector

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

which allow us to write the model simply as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

- In multiple regression where we have k explanatory variables plus the intercept
$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times(k+1)}{\mathbf{X}} \underset{(k+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{u}}$$
where
$$\underset{n\times(k+1)}{\mathbf{X}} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$
and
$$\underset{(k+1)\times 1}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$
- Remember that $\mathbf{y}, \mathbf{X}$ are observable, $\boldsymbol{\beta}$ and $\mathbf{u}$ are unknown and unobservable
- Also remember that $\mathbf{Xb}$ for any $(k+1)\times 1$ vector $\mathbf{b}$ is an $n \times 1$ vector that is a linear combination of columns of $\mathbf{X}$

The power of abstraction

- In the real world we have goals such as:
  - We want to determine the added value of education to wage after controlling for experience and IQ based on a random sample of 1000 observations
  - We want to establish if the return on a firm's share price is related to its value measured by its book to market ratio and firm's size based on 24 yearly observations on 500 stocks
  - We want to predict the value of house in the city of Monash given characteristics such as land area, number of bedrooms, number of bathrooms, proximity to public transport, etc., based on data on the last 100 houses sold in Monash
  - We want to forecast hourly electricity load based on all information available at the time that forecast is made, based on hourly electricity load in the last 3 years
- In all cases we want to get a good estimate of $\boldsymbol{\beta}$ in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ based on our sample of observations
- We have now turned these very diverse questions into a single mathematical problem

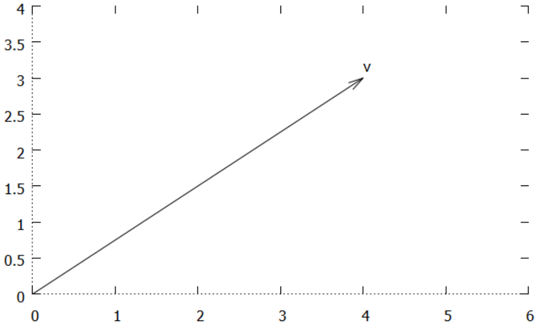
The OLS solution

- OLS finds the vector in the column space of $\mathbf{X}$ which is closest to $\mathbf{y}$. Let's see what this means
- To visualise we are limited to 3 dimensions at most
- Assume we want to explain house prices with number of bedrooms
- We have data on 3 houses, with 4, 1 and 1 bedrooms which sold for 10, 4 and 6 $\times\$100,000$
- Want to find a good $\widehat{\boldsymbol{\beta}}$

$$\begin{bmatrix} 10 \\ 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \hat{\beta}_0 + \begin{bmatrix} 4 \\ 1 \\ 1 \end{bmatrix} \hat{\beta}_1 + \hat{\mathbf{u}} = \begin{bmatrix} 1 & 4 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \widehat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$$

Vectors and vector spaces

- An n-dimensional vector is an arrow from the origin to the point in $\Re^n$ with coordinates given by its elements
- Example: $\mathbf{v} = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$



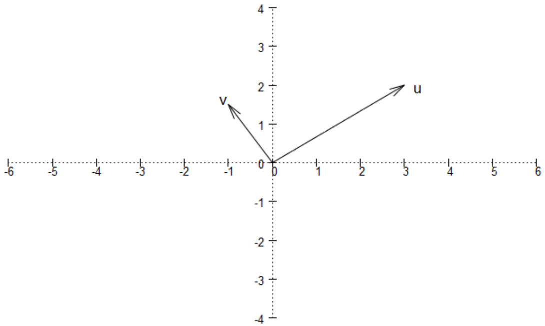
- The **length** of a vector is the square root of sum of squares of its coordinates

$$length(\mathbf{v}) = (\mathbf{v}'\mathbf{v})^{1/2}$$

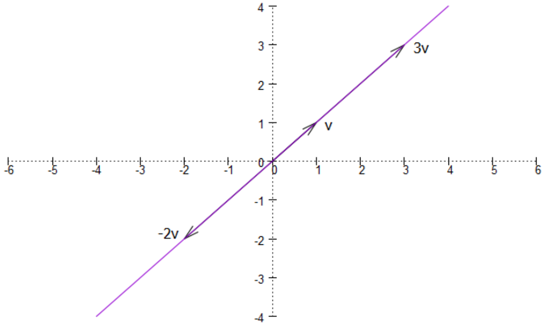
- Example: vector **v** on the previous slide
- For **u** and **v** of the same dimension

$$\mathbf{u}'\mathbf{v} = 0 \Leftrightarrow \mathbf{u} \text{ and } \mathbf{v} \text{ are perpendicular (orthogonal) to each other}$$

- Example: $\begin{bmatrix} 3 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 1.5 \end{bmatrix}$



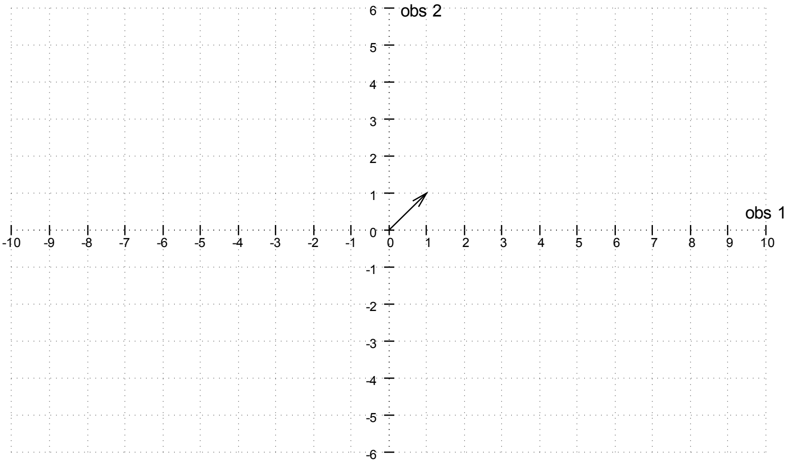
- For any constant c, the vector $c\mathbf{v}$ is on the line that passes through $\mathbf{v}$
- This line is called the "space spanned by $\mathbf{v}$"
- Example: $\begin{bmatrix} 3 \\ 3 \end{bmatrix}$ and $\begin{bmatrix} -2 \\ -2 \end{bmatrix}$ are in the space spanned by $\mathbf{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

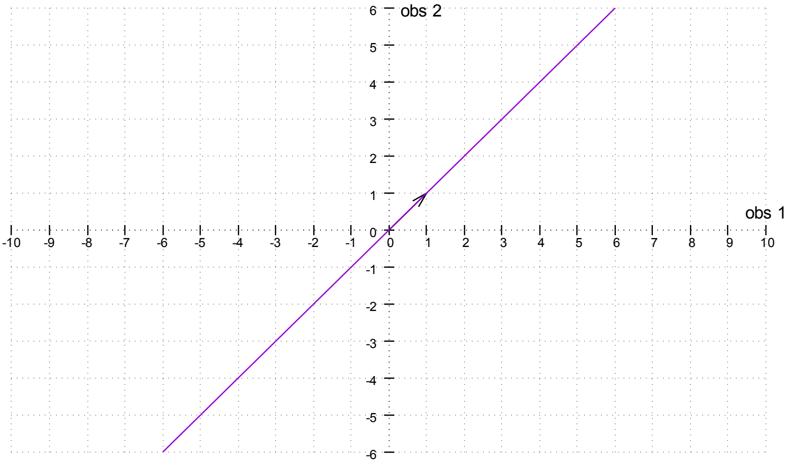


- Consider a matrix $\mathbf{X}$ that has two columns that are not a multiple of each other
- Geometrically, these vectors form a two dimensional plane that contains all linear combinations of these two vectors, i.e. set of all $\mathbf{Xb}$ for all $\mathbf{b}$. This plane is called the column space of $\mathbf{X}$
- If the number of rows of $\mathbf{X}$ is also two, then the column space of $\mathbf{X}$ is the entire $\Re^2$, that is any two dimensional $\mathbf{y}$ can be written as a linear combination of columns of $\mathbf{X}$
- Consider the house price example, but only with two houses

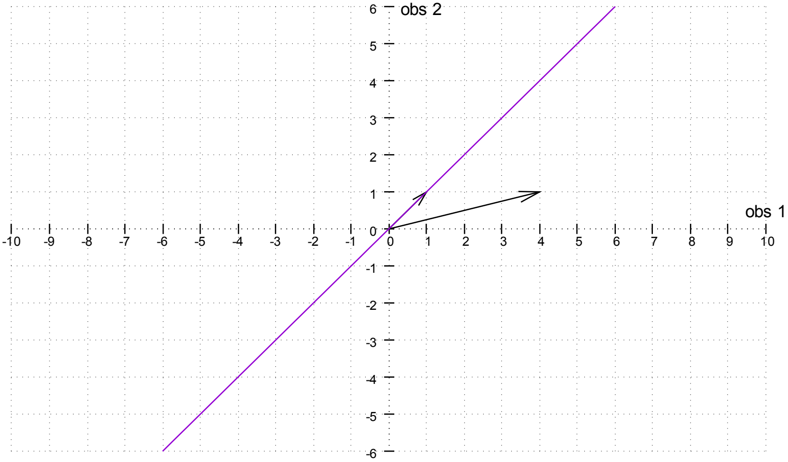
$$\begin{bmatrix} 10 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \hat{\beta}_0 + \begin{bmatrix} 4 \\ 1 \end{bmatrix} \hat{\beta}_1 + \hat{\mathbf{u}} = \begin{bmatrix} 1 & 4 \\ 1 & 1 \end{bmatrix} \widehat{\boldsymbol{\beta}} + \hat{\mathbf{u}}$$

- The price vector can be perfectly explained with a linear combination of columns of $\mathbf{X}$, i.e. with zero $\hat{\mathbf{u}}$

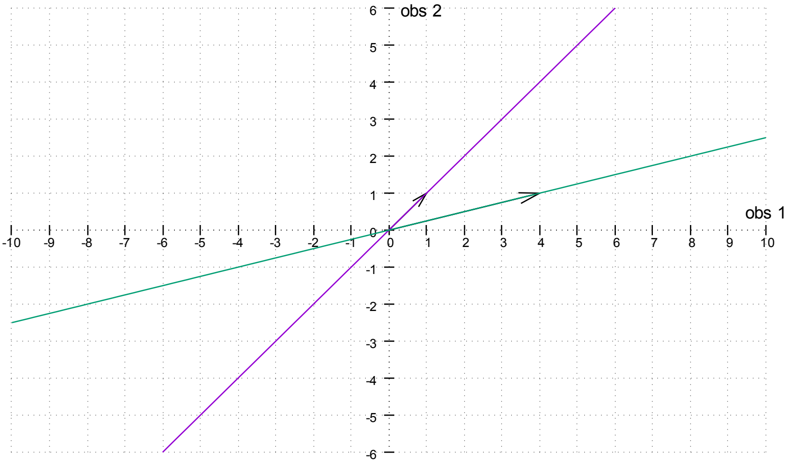

obs 2
obs 1
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10

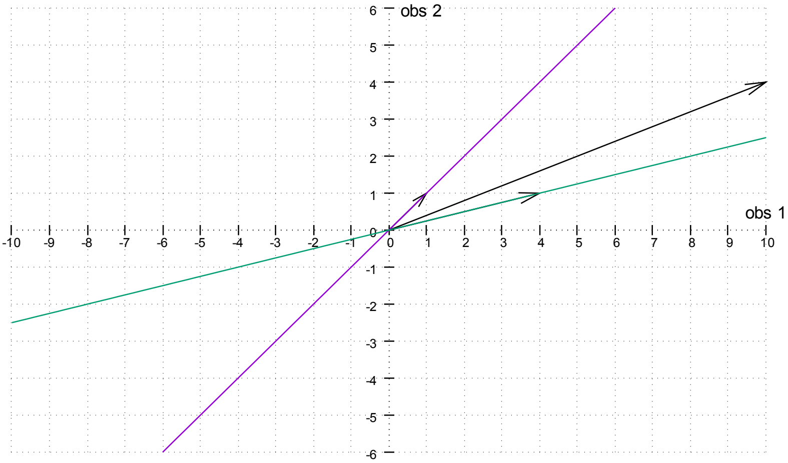


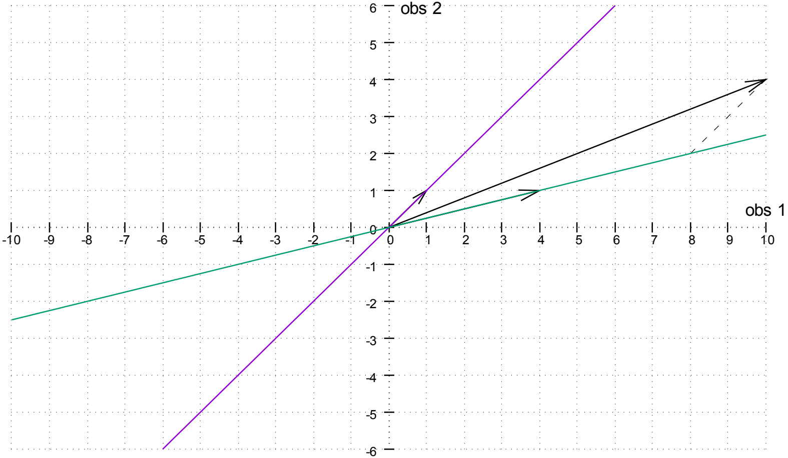
obs 2
obs 1
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10

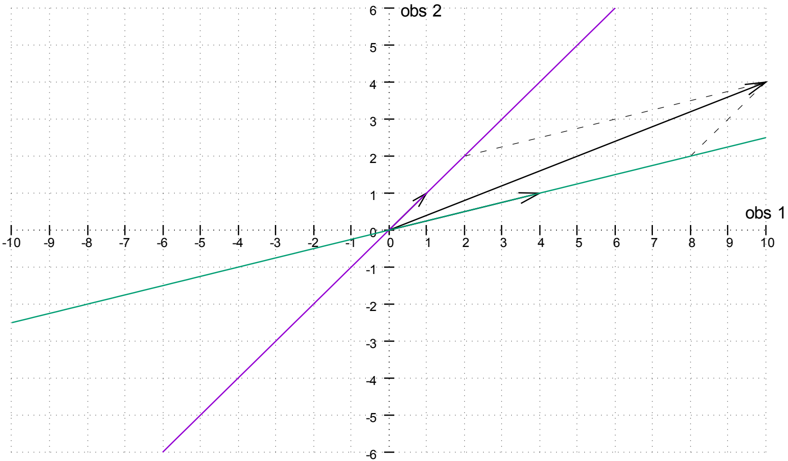


obs 2
obs 1
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10


obs 2
obs 1
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10


obs 2
obs 1
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10

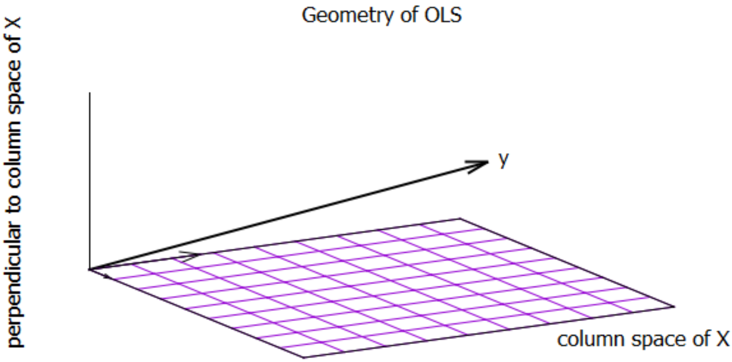

obs 2
obs 1
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
1
2
3
4
5
6
7
8
9
10


obs 2
obs 1
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-10
-9
-8
-7
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10

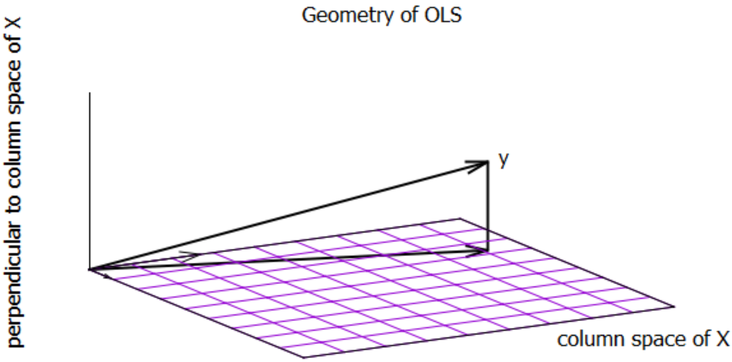
- This shows that with only two observations, we would suggest

$$\widehat{price} = 2 + 2bedrooms$$

- While this fits the first two houses perfectly, when we add the third house (1 bedroom sold for 6 hundred thousands) we make an error of 2 hundred thousand dollars
- With 3 observations, the 3 dimensional price vector no longer lies in the space of linear combinations of columns of $\mathbf{X}$.
- The closest point in the column space of $\mathbf{X}$ to $\mathbf{y}$ is ...



Geometry of OLS
perpendicular to column space of X
y
column space of X


Geometry of OLS
perpendicular to column space of X
y
column space of X

- Hence, the shortest $\hat{\mathbf{u}}$ is the one that is perpendicular to column of $\mathbf{X}$, i.e.

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

- Since $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$, we get

$$\begin{aligned}
\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) &= 0 \\
\Rightarrow \mathbf{X}'\mathbf{y} &= \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} \\
\Rightarrow &\boxed{\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}
\end{aligned}$$

- For $\mathbf{X}'\mathbf{X}$ to be invertible, columns of $\mathbf{X}$ must be *linearly independent*
- The top box is more important than the formula for the OLS estimator in the second box. Geometry of OLS is summarised in the first box: The OLS residuals are orthogonal to columns of $\mathbf{X}$

- The vector of OLS predicted values is the orthogonal projection of $\mathbf{y}$ in the column space of $\mathbf{X}$

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

- By definition

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{u}}$$

- Since $\mathbf{y}$, $\hat{\mathbf{y}}$ and $\hat{\mathbf{u}}$ form a right-angled triangle, we know (remember the length of a vector)

$$\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\mathbf{u}}'\hat{\mathbf{u}}$$

i.e.,

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} \hat{u}_i^2 \tag{L2}$$

- Since $(1 \quad 1 \quad \cdots \quad 1)\hat{\mathbf{u}} = 0$ we also have

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \Rightarrow \bar{y} = \bar{\hat{y}}$$

- Subtracting $n\bar{y}^2$ from both sides of (L2) and rearranging, we get

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

or

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

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

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

Interpretation of OLS estimates

Example: The causal effect of education on wage

- Consider an extension of the *wage* equation that is in tutorial 2:

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

  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.

OLS in action

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 hold x_2 fixed, that is, its change is zero, $\Delta x_2 = 0$?

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

- In particular,

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

- In other words, $\hat{\beta}_1$ is the slope of $\hat{y}$ with respect to x_1 when 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,

$$\Delta \hat{y} = \hat{\beta}_2 \Delta x_2 \text{ if } \Delta x_1 = 0$$

and

$$\hat{\beta}_2 = \frac{\Delta \hat{y}}{\Delta x_2} \text{ if } \Delta x_1 = 0$$

- Let's go back to regression output and interpret the parameters.

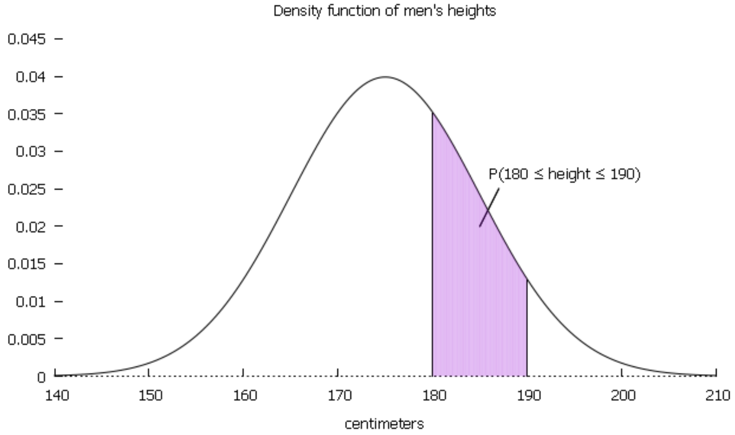
$$\begin{aligned} \widehat{wage} &= -128.89 + 42.06\ educ + 5.14\ IQ \\ n &= 935,\ R^2 = .134 \end{aligned}$$

- 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*")

Summary

- Given a sample of n observations, OLS finds the orthogonal projection of the dependent variable vector in the column space of explanatory variables
- The residual vector is orthogonal to each of the explanatory variable vectors, including a column of ones for the intercept
$$\mathbf{X}'\hat{\mathbf{u}} = 0$$
- This leads to the formula for the OLS estimator in multiple regression
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$
- It also leads to
$$\text{SST} = \text{SSE} + \text{SSR}$$
- We learned how to interpret the coefficients of an estimated regression model
- Note the difference between the population parameter $\boldsymbol{\beta}$ and its OLS estimator $\widehat{\boldsymbol{\beta}}$. $\boldsymbol{\beta}$ is constant and does not change, $\widehat{\boldsymbol{\beta}}$ is a function of sample and its value changes for different samples. Why are these good estimators?

- The probability that the height of a randomly selected man lies in a certain interval is the area under the pdf over that interval



Back to what is a model