# UGA M1: Econometrics 1
# Introduction to Regression

Michal W. Urdanivia*

*Université de Grenoble Alpes, Faculté d'Économie, GAEL,
e-mail: michal.wong-urdanivia@univ-grenoble-alpes.fr

September 25, 2017

# Part I

# Definition and interpretation of regression

1. Motivation

2. Conditional expectation function

3. Population regression

3.1 Interpretation

4. Sample regression

5. Regression in R

# References

- Angrist and Pischke (2014) chapter 2
- Wooldridge (2013) chapter 2
- Stock and Watson (2009) chapter 4-5
- Angrist and Pischke (2009) chapter 3 up to and including section 3.1.2 (pages 27-40)
- Abbring (2001) chapter 3
- Diez, Barr, and Cetinkaya-Rundel (2012) chapter 7
- Bierens (2012)
- Baltagi (2002) chapter 3

# References

- The most useful reference is likely Wooldridge (2013) or Stock and Watson (2009), followed by Angrist and Pischke (2014).
- Angrist and Pischke (2009) is also very nice, but a bit more dense. Diez, Barr, and Cetinkaya-Rundel (2012) is a simple introduction to regression with many examples and not much math.
- Baltagi (2002) is more technical and difficult than Wooldridge, but I think would still be useful.
- Bierens (2012) has many of the proofs that we will go through, but the typesetting is not great.
- Abbring (2001) moves quickly and uses matrix notation.

Section 1

# Motivation

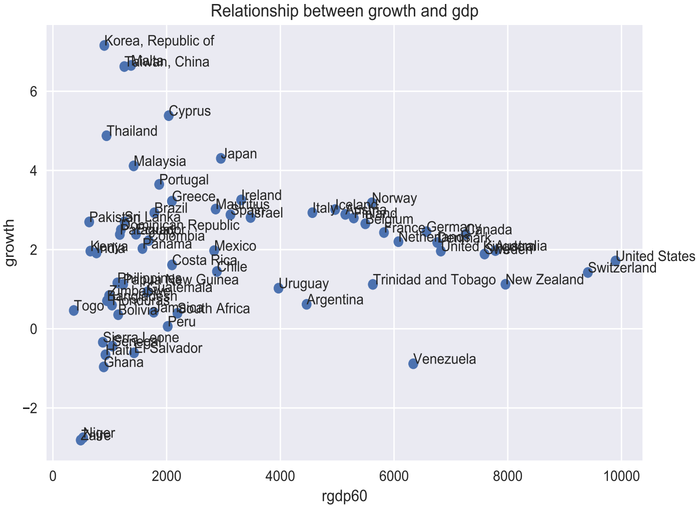
## General problem

- Often interested in relationship between two (or more) variables, e.g.
  - Wages and education
  - Minimum wage and unemployment
  - Price, quantity, and product characterics
- Usually have:
  1. Variable to be explained or response variable or outcome or regressand or dependent variable.
  2. Regressors or covariates or explanatory variable(s) or independent variables.
- A usual notation for the response variable is $Y$ and for the covariates $X$.
- Example:

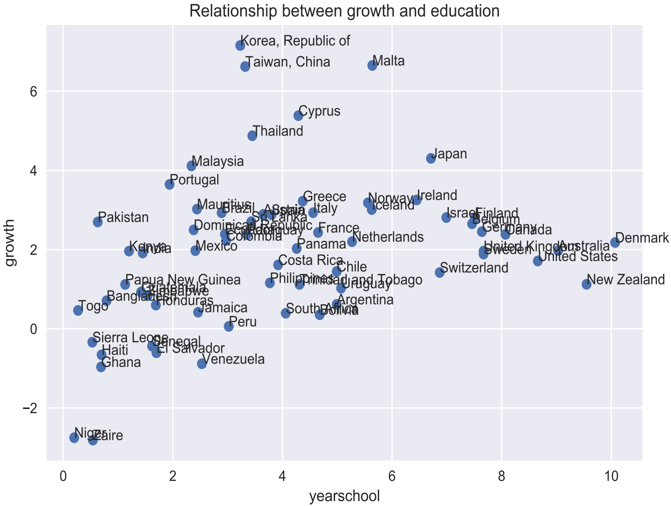
| **Response** | **Covariates** |
|---|---|
| Wage | Education |
| Unemployment | Minimum wage |
| Quantity | Price and product characteristics |

- For now agnostic about causality, but $\mathbb{E}[Y|X]$ usually is not causal

# Example: Growth and GDP



# Example: Education and GDP



## Conditional expectation function

- One way to describe relation between two variables is a function,

$$Y = h(X)$$

- Most relationships in data are not deterministic, so look at average relationship,

$$\begin{aligned} Y &= \underbrace{\mathbb{E}[Y|X]}_{\equiv h(X)} + \underbrace{(Y - \mathbb{E}[Y|X])}_{\equiv U} \\ &= \mathbb{E}[Y|X] + U \end{aligned}$$

- Note that $\mathbb{E}[U] = 0$ (by definition of $U$ and iterated expectations)
- $\mathbb{E}[Y|X]$ can be any function, in particular, it need not be linear
- Unrestricted $\mathbb{E}[Y|X]$ hard to work with
  - Hard to estimate
  - Hard to communicate if $X$ a vector (cannot draw graphs)
- Instead use linear regression
  - Easier to estimate and communicate
  - Tight connection to $\mathbb{E}[Y|X]$

# Population regression

- The bivariate population regression of $Y$ on $X$ is

$$(\beta_0, \beta_1) = \underset{b_0, b_1}{\arg\min}\, \mathbb{E}[(Y - b_0 - b_1 X)^2]$$

  i.e. $\beta_0$ and $\beta_1$ are the slope and intercept that minimize the expected square error of $Y - (\beta_0 + \beta_1 X)$

- Calculating $\beta_0$ and $\beta_1$:
  - First order conditions:

$$\begin{aligned}[b_0] : 0 &= \frac{\partial}{\partial b_0} \mathbb{E}[(Y - b_0 - b_1 X)^2] \\ &= \mathbb{E}\left[\frac{\partial}{\partial b_0}(Y - b_0 - b_1 X)^2\right] \\ &= \mathbb{E}\left[-2(Y - \beta_0 - \beta_1 X)\right] \qquad (1)\end{aligned}$$

# Population regression

and

$$
\begin{aligned}
[b_1] : 0 &= \frac{\partial}{\partial b_1} \mathbb{E}[(Y - b_0 - b_1 X)^2] \\
&= \mathbb{E}\left[\frac{\partial}{\partial b_1}(Y - b_0 - b_1 X)^2\right] \\
&= \mathbb{E}\left[-2(Y - \beta_0 - \beta_1 X)X\right] \qquad (2)
\end{aligned}
$$

- (1) rearranged gives $\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X]$
- Substituting into (2)

$$
\begin{aligned}
0 &= \mathbb{E}\left[X(-Y + \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] + \beta_1 X)\right] \\
&= \mathbb{E}\left[X(-Y + \mathbb{E}[Y])\right] + \beta_1 \mathbb{E}\left[X(X - \mathbb{E}[X])\right] \\
&= -\mathbb{C}(X, Y) + \beta_1 \mathbb{V}(X) \\
\beta_1 &= \frac{\mathbb{C}(X, Y)}{\mathbb{V}(X)}
\end{aligned}
$$

- $\beta_1 = \frac{\mathbb{C}(X,Y)}{\mathbb{V}(X)}$, $\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X]$

# Population regression approximates $\mathbb{E}[Y|X]$

## Lemma 1

*The population regression is the minimal mean square error linear approximation to the conditional expectation function, i.e.*

$$\underbrace{\arg\min_{b_0, b_1} \mathbb{E}\left[(Y-(b_0+b_1X))^2\right]}_{\textit{population regression}} = \arg\min_{b_0, b_1} \underbrace{\mathbb{E}_X\left[(\mathbb{E}[Y|X]-(b_0+b_1X))^2\right]}_{\textit{MSE of linear approximation to } \mathbb{E}[Y|X]}$$

## Corollary 2

*If* $\mathbb{E}[Y|X] = c + mX$, *then the population regression of* $Y$ *on* $X$ *equals* $\mathbb{E}[Y|X]$, *i.e.* $\beta_0 = c$ *and* $\beta_1 = m$

# Proof

## Proof.

- Let $b_0^*, b_1^*$ be minimizers of MSE of approximation to $\mathbb{E}[Y|X]$
- Same steps as in population regression formula gives

$$0 = \mathbb{E}\left[-2(\mathbb{E}[Y|X] - b_0^* - b_1^* X)\right]$$

and

$$0 = \mathbb{E}\left[-2(\mathbb{E}[Y|X] - b_0^* - b_1^* X)X\right]$$

- Rearranging and combining,

$$b_0^* = \mathbb{E}[\mathbb{E}[Y|X]] - b_1^*\mathbb{E}[X] = \mathbb{E}[Y] - b_1^*\mathbb{E}[X]$$

and

$$\begin{aligned}
0 =& \mathbb{E}\left[X(-\mathbb{E}[Y|X] + \mathbb{E}[Y] + b_1^*\mathbb{E}[X] - b_1^* X)\right] \\
=& \mathbb{E}\left[X(-\mathbb{E}[Y|X] + \mathbb{E}[Y])\right] + b_1^*\mathbb{E}\left[X(X - \mathbb{E}[X])\right] \\
=& -\mathbb{C}(X, Y) + b_1^*\mathbb{V}(X) \\
b_1^* =& \frac{\mathbb{C}(X, Y)}{\mathbb{V}(X)}
\end{aligned}$$

□

## Regression interpretation

- Regression = best linear approximation to $\mathbb{E}[Y|X]$
- $\beta_0 \approx \mathbb{E}[Y|X=0]$
- $\beta_1 \approx \frac{d}{dx}\mathbb{E}[Y|X] \approx$ change in average $Y$ per unit change in $X$
- Not necessarily a causal relationship (usually not)
- Always can be viewed as description of data

## Regression with binary $X$

- Suppose $X$ is binary (i.e. can only be 0 or 1)
- We know $\beta_0 + \beta_1 X =$ best linear approximation to $\mathbb{E}[Y|X]$
- $X$ only takes two values,
  - $\beta_0 = \mathbb{E}[Y|X = 0]$
  - $\beta_0 + \beta_1 = \mathbb{E}[Y|X = 1]$

# Sample regression

- Have sample of observations: $\{(Y_i, X_i)\}_{i=1}^{N}$
- The sample regression (or when unambiguous just "regression") of $Y$ on $X$ is

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{b_0, b_1}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2$$

i.e. $\hat{\beta}_0$ and $\hat{\beta}_1$ are the slope and intercept that minimize the sum of squared errors, $(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$

  - Same as population regression but with sample average instead of expectation

- Same calculation as for population regression would show

$$\hat{\beta}_1 = \frac{\widehat{\mathbb{C}}(X, Y)}{\widehat{\mathbb{V}(X)}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{N})(Y_i - \bar{Y})}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

and

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

## Sample regression

- Since $\hat{\beta}_1$ and $\hat{\beta}_0$ come from minimizing a sum of squares, they are called the ordinary least squares estimates, or OLS for short.
- The formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$ come from the first order conditions.
- These estimators minimize the sum of squared differences between the regression line and the observed $Y_i$,

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{b_0, b_1}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2.$$

The first order condition for $\hat{\beta}_0$ is:

$$0 = \frac{1}{N} \sum_{i=1}^{N} 2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

# Sample regression

which can be rearranged to get

$$\hat{\beta}_0 = \frac{1}{N}\left(\sum_{i=1}^{N} Y_i - \hat{\beta}_1 X_i\right)$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{N}.$$

The first order condition for $\hat{\beta}_1$ is:

$$0 = \frac{1}{N}\sum_{i=1}^{N} 2X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i).$$

Substituting in the previous expression for $\hat{\beta}_0$ gives

$$0 = \frac{1}{N}\sum_{i=1}^{N} 2X_i(Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i).$$

## Sample regression

Rearranging to solve for $\hat{\beta}_1$:

$$\hat{\beta}_1 \sum_{i=1}^{N} X_i(X_i - \bar{X}) = \sum_{i=1}^{N} X_i(Y_i - \bar{Y})$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} X_i(Y_i - \bar{Y})}{X_i(X_i - \bar{X})} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{(X_i - \bar{X})^2}.$$

# Sample regression

- Sample regression is an estimator for the population regression
- Given an estimator we should ask:
    - Unbiased?
    - Variance?
    - Consistent?
    - Asymptotically normal?
- We will address these questions in the next week or two

# True linear model approach to regression

- Most authors (like Wooldridge) introduce regression by starting with a linear model for $Y$:

$$Y = \alpha_0 + \alpha_1 X + \underbrace{U}_{\text{unobserved}}$$

with $\mathbb{E}[UX] = 0$

  - Perspective: model = true description of data generating process
- Implications:
  - Population regression coefficients = model coefficients, i.e. $\beta_0 = \alpha_0$ and $\beta_1 = \alpha_1$
  - Conditional expectation function is linear

$$\mathbb{E}[Y|X] = \alpha_0 + \alpha_1 X$$

  - $\beta_1 = \alpha_1$ has causal interpretation as long as we believe $\mathbb{E}[UX] = 0$
  - Easier to discuss causality
  - Easier to derive statistical properties

# Problems with true linear models

- Usually do not believe models are linear, e.g. no economic theory that says the following should be linear:
  1. $\log(wage_i) = \beta_0 + \beta_1(educ_i) + U_i$
  2. $\log q_t = \beta_0 + \beta_1 \log p_t + U_t$
- Usually do not believe error terms are uncorrelated with covariates, e.g.
  1. Need $\mathbb{E}[U_i educ_i] = 0$, but $U_i$ probably includes IQ, propensity to work hard, etc, which should be correlated with education
  2. Is it supposed to be demand or supply? Either case, changes in $q_t$ from $U_t$ generally also change equilibrium $p_t$, so $\mathbb{E}[\log p_t U_t] \neq 0$
- Viewing regression as best linear approximation to $\mathbb{E}[Y|X]$ makes it clear what regression tells you about the data even if the true model is not linear and does not have $\mathbb{E}[XU] = 0$

# Part II

## Properties of regression

6. Fitted value and residuals

7. Statistical properties

7.1 Unbiased

7.2 Variance

7.3 Distribution

7.4 Discussion of assumptions

8. Examples

## Fitted values and residuals

- These algebraic identities about fitted values and residuals are things that we will use repeatedly later.
- I would not recommend spending time trying to memorize these.
- The important ones will come up repeatedly and you will remember them without any special effort.
- The first time we use these identities, we will go through how to get them again. We may even go through them yet again the second and third time we use them.
- Eventually we will use some of these identities so often that you will either be able to quickly derive them or just remember them.
- Fitted values:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- Residuals:

$$\hat{U}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = Y_i - \hat{Y}_i$$

$$Y_i = \hat{Y}_i + \hat{U}_i$$

# Fitted values and residuals

- Sample mean of residuals $= 0$
  - First order condition for $\hat{\beta}_0$,

$$0 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)$$

$$0 = \frac{1}{N}\sum_{i=1}^{N}\hat{U}_i$$

- Sample covariance of $X$ and $\hat{U} = 0$
  - First order condition for $\hat{\beta}_1$,

$$0 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i$$

$$0 = \frac{1}{N}\sum_{i=1}^{N}\hat{U}_i X_i$$

# Fitted values and residuals

- Sample mean of $\hat{Y}_i = \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} Y_i =& \frac{1}{N}\sum_{i=1}^{N} \hat{Y}_i + \hat{U}_i \\
=& \frac{1}{N}\sum_{i=1}^{N} \hat{Y}_i \\
=& \frac{1}{N}\sum_{i=1}^{N} \hat{\beta}_0 + \hat{\beta}_1 X_i \\
=& \hat{\beta}_0 + \hat{\beta}_1 \bar{X}
\end{aligned}$$

## Fitted values and residuals

- Sample covariance of $Y$ and $\hat{U}$ = sample variance of $\hat{U}$:

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} Y_i(\hat{U}_i - \bar{\hat{U}}) =& \frac{1}{N}\sum_{i=1}^{N} Y_i \hat{U}_i \\
=& \frac{1}{N}\sum_{i=1}^{N} (\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{U}_i)\hat{U}_i \\
=& \hat{\beta}_0 \frac{1}{N}\sum_{i=1}^{N} \hat{U}_i + \beta_1 \frac{1}{N}\sum_{i=1}^{N} X_i \hat{U}_i + \frac{1}{N}\sum_{i=1}^{N} \hat{U}_i^2 \\
=& \frac{1}{N}\sum_{i=1}^{N} \hat{U}_i^2
\end{aligned}$$

## $R^2$

- Decompose $Y_i$

$$Y_i = \hat{Y}_i + \hat{U}_i$$

- Total sum of squares = explained sum of squares + sum of squared residuals

$$\underbrace{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}_{SST} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}_{SSE} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\hat{U}_i^2}_{SSR}$$

- R-squared: fraction of sample variation in $Y$ that is explained by $X$

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = o\widehat{\mathrm{Corr}}(Y, \hat{Y})$$

  - $0 \leq R^2 \leq 1$
  - If all data on regression line, then $R^2 = 1$
  - Magnitude of $R^2$ does not have direct bearing on economic importance of a regression

## Unbiased

- $\mathbb{E}[\hat{\beta}] = ?$
- Assume:

SLR.1 (linear model) $Y_i = \beta_0 + \beta_1 X_i + U_i$

SLR.2 (independence) $\{(X_i, Y_i)\}_{i=1}^{N}$ is independent random sample

SLR.3 (rank condition) $\widehat{\mathbb{V}}(X) > 0$

SLR.4 (exogeneity) $\mathbb{E}[U|X] = 0$

- Then, $\mathbb{E}[\hat{\beta}_1] = \beta_1$ and $\mathbb{E}[\hat{\beta}_0] = \beta_0$
- It is more important to understand the meaning of these four assumptions (discussed below) than the proof that regression is unbiased.

Regression is unbiased.

We need to calculate $\mathbb{E}[\hat{\beta}]$. First, substitute in the formula for $\hat{\beta}$.

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{\sum_{i=1}^{N}(X_i-\bar{X})Y_i}{\sum_{i=1}^{N}(X_i-\bar{X})X}\right]$$

Next, substitute in the model for $Y_i$, $Y_i = \beta_0 + \beta_1 X_i + U_i$,

$$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\frac{\sum_{i=1}^{N}(X_i-\bar{X})(\beta_0+\beta_1 X_i+U_i)}{\sum_{i=1}^{N}(X_i-\bar{X})X}\right]$$

rearrange

$$= \mathbb{E}\left[\frac{\overbrace{\sum_{i=1}^{N} X_i - \bar{X}}^{=0}}{\sum_{i=1}^{N}(X_i-\bar{X})X}\beta_0 + \overbrace{\left(\frac{\sum_{i=1}^{N}(X_i-\bar{X})X_i}{\sum_{i=1}^{N}(X_i-\bar{X})X}\right)}^{=1}\beta_1 + \frac{\sum_{i=1}^{N}(X_i-\bar{X})U_i}{\sum_{i=1}^{N}(X_i-\bar{X})X}\right]$$

use linearity of expectation

$$= \beta_1 + \mathbb{E}\left[\frac{\sum_{i=1}^{N}(X_i-\bar{x})U_i}{\sum_{i=1}^{N}(X_i-\bar{X})X}\right]$$

use iterated expectations

$$= \beta_1 + \mathbb{E}_X\left[\mathbb{E}_{U|X}\left[\frac{\sum_{i=1}^{N}(X_i-\bar{X})U_i}{\sum_{i=1}^{N}(X_i-\bar{X})X}\middle| X_1, X_2, ..., X_n\right]\right]$$

conditional on $X_1, ..., X_n$, $X_i$ is constant

$$= \beta_1 + \mathbb{E}_X\left[\frac{\sum_{i=1}^{N}(X_i-\bar{X})\mathbb{E}[U_i|X_1, ..., X_n]}{\sum_{i=1}^{N}(X_i-\bar{X})X}\right]$$

independent observations implies $\mathbb{E}[U_i|X_1, ..., X_n] = \mathbb{E}[U_i|X_i]$

$$= \beta_1 + \mathbb{E}_X\left[\frac{\sum_{i=1}^{N}(X_i-\bar{X})\mathbb{E}[U_i|X_i]}{\sum_{i=1}^{N}(X_i-\bar{X})X}\right]$$

exogeneity assumption says that $\mathbb{E}[U_i|X_i] = 0$.

$$= \beta_1$$

□

- Note that the first few steps of the above proof showed that

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N}(X_i - \bar{X})U_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2}.$$

- This is a very useful expression that can be used as a starting point for calculating the variance of $\hat{\beta}_1$, and thinking about what happens if exogeneity fails and $\mathbb{E}[U|X] \neq 0$.

# Variance

- $\mathbb{V}(\hat{\beta})$?
- Assume SLR.1-4 and

SLR.5 (homoskedasticity) $\mathbb{V}(U|X) = \sigma^2$

- Then,

$$\mathbb{V}(\hat{\beta}_1|\{X_i\}_{i=1}^n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

and

$$\mathbb{V}(\hat{\beta}_0|\{X_i\}_{i=1}^N) = \frac{\sigma^2 \frac{1}{N} \sum\limits_{i=1}^N X_i^2}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

# Variance

- As in the proof that regression is unbiased,

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N}(X_i - \bar{X})U_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2}.$$

- We now want to take the variance of this expression.
- Before doing so, it will be useful to review some properties of the variance of a sum of random variables.

## Lemma 3

*Let a, b, and c be constants, and Z and W be random variables. Then,*

$$\mathbb{V}(a + bZ + cW) = b^2\mathbb{V}(Z) + c^2\mathbb{V}(W) + 2bc\mathbb{C}(Z, W).$$

- We can prove this using the definition of variance.

Proof.

$$\begin{aligned}
\mathbb{V}(a+bZ+cW) &= \mathbb{E}\left[(a+bZ+cW-\mathbb{E}[a+bZ+cW])^2\right] \\
&= \mathbb{E}\left[(a+bZ+cW-a-b\mathbb{E}[Z]-c\mathbb{E}[W])^2\right] \\
&= \mathbb{E}\left[(b(Z-\mathbb{E}[Z])+c(W-\mathbb{E}[W]))^2\right] \\
&= \mathbb{E}\left[b^2(Z-\mathbb{E}[Z])^2+2bc(Z-\mathbb{E}[Z])(W-\mathbb{E}[W])\right. \\
&\quad \left.+\ c^2(W-\mathbb{E}[W])^2\right] \\
&= b^2\mathbb{E}\left[(Z-\mathbb{E}[Z])^2\right]+2bc\mathbb{E}\left[(Z-\mathbb{E}[Z])(W-\mathbb{E}[W])\right] \\
&\quad +c^2\mathbb{E}\left[(W-\mathbb{E}[W])^2\right] \\
&= b^2\mathbb{V}(Z)+c^2\mathbb{V}(W)+2bc\mathbb{C}(Z,W)
\end{aligned}$$

□

- Generalizing the above to the sum of more than two random variables, we have

## Corollary 4

*Let $a_1, ..., a_n$ be constants, and $Z_1, ..., Z_n$ be random variables, then,*

$$\mathbb{V}\left(\sum_{i=1}^{N} a_i Z_i\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j \mathbb{C}(Z_i, Z_j)$$

*Furthermore, if $Z_i$ and $Z_j$ are independent (or just uncorrelated) for $i \neq j$, then*

$$\mathbb{V}\left(\sum_{i=1}^{N} a_i Z_i\right) = \sum_{i=1}^{N} a_i^2 \mathbb{V}(Z_i)$$

- We can apply this corollary to

$$\begin{aligned}
\mathbb{V}(\hat{\beta}_1|x) &= \mathbb{V}\left(\beta_1 + \frac{\sum_{i=1}^N (X_i-\bar{X})U_i}{\sum_{i=1}^N (X_i-\bar{X})^2}\middle| x\right) \\
&= \mathbb{V}\left(\sum_{i=1}^N \underbrace{\frac{X_i-\bar{X}}{\sum_{i=1}^N (X_i-\bar{X})^2}}_{a_i}\underbrace{U_i}_{Z_i}\middle| x\right) \qquad \text{using the corollary} \\
&= \sum_{i=1}^N\sum_{j=1}^N \frac{X_i-\bar{X}}{\sum_{i=1}^N (X_i-\bar{X})^2}\frac{x_j-\bar{X}}{\sum_{i=1}^N (X_i-\bar{X})^2}\mathbb{C}(U_i,U_j|X) \qquad \text{independence} \\
&= \sum_{i=1}^N \left(\frac{X_i-\bar{X}}{\sum_{i=1}^N X_i-\bar{X}}\right)^2 \mathbb{V}(U_i|X) \qquad \text{homoskedasticity} \\
&= \sum_{i=1}^N \frac{(X_i-\bar{X})^2}{\left(\sum_{i=1}^N (X_i-\bar{X})^2\right)^2}\sigma_U^2 \\
&= \frac{\sigma_U^2}{\sum_{i=1}^N (X_i-\bar{X})^2}.
\end{aligned}$$

## Distribution with normal errors

- Assume SLR.1-SLR.5 and

SLR.6 (normality) $U_i|X_i \sim N(0, \sigma^2)$

- Then $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$, and

$$\hat{\beta}_1|\{X_i\}_{i=1}^N \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^N (X_i - \bar{X})^2}\right)$$

- Even without assuming normality, the central limit theorem implies $\hat{\beta}$ is asymptotically normal (details in a later lecture)

# Distribution with normal errors

- An important property of normal random variables is that if $Z$ and $W$ are independent, and $Z \sim N(\mu_z, \sigma_z^2)$ and $W \sim N(\mu_w, \sigma_w^2)$, then

$$a + bZ + cW \sim N(a + b\mu_z + c\mu_w, b^2\sigma_z^2 + c^2\sigma_w^2).$$

- If we assume that $U_i$ is normally distributed conditional on $X$, then since $\hat{\beta}_1$ is just a sum of the $U_i$,[1] $\hat{\beta}_1$ will also be normally distributed.

[1] Specifically, $\hat{\beta}_{\mathbf{1}} = \beta_{\mathbf{1}} + \frac{\sum_{i=\mathbf{1}}^{N}(X_i - \bar{X})U_i}{\sum_{i=\mathbf{1}}^{N}(X_i - \bar{X})^{\mathbf{2}}}$.

# Summary

- Simple linear regression model assumptions:

SLR.1 (linear model) $Y_i = \beta_0 + \beta_1 X_i + U_i$
SLR.2 (independence) $\{(X_i, Y_i)\}_{i=1}^{n}$ is independent random sample
SLR.3 (rank condition) $\widehat{\mathbb{V}}(X) > 0$
SLR.4 (exogeneity) $\mathbb{E}[U|X] = 0$
SLR.5 (homoskedasticity) $\mathbb{V}(U|X) = \sigma^2$
SLR.6 (normality) $U_i|X_i \sim N(0, \sigma^2)$

- $\hat{\beta}$ unbiased if SLR.1-SLR.4
- If also SLR.5, then $\mathbb{V}(\hat{\beta}_1|\{X_i\}_{i=1}^{N}) = \frac{\sigma^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$
- If also SLR.6, then $\hat{\beta}_1|\{X_i\}_{i=1}^{N} \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right)$

# Discussion of assumptions

SLR.1 Having a linear model makes it easier to state the other assumptions, but we could instead start by saying let $\beta_1 = \frac{\mathbb{C}(X,Y)}{\mathbb{V}(X)}$ and $\beta_0 = \mathbb{E}[Y] - \beta_1\mathbb{E}[X]$ be the population regression coefficients and define $U_i = Y_i - \beta_0 - \beta_1 X_i$

# Discussion of assumptions

- To say whether an estimator is unbiased, we first have to define what parameter we want to estimate.
- Assuming that there is a linear model defines the parameter we want to estimate.
- This linear model could be population regression, in which case, $\beta_1 = \frac{\mathbb{C}(X,Y)}{\mathbb{V}(X)}$, and by construction we must have $\mathbb{E}[XU] = 0$.
- However, the linear model may also be motivated by economic theory.
- For example, consider a Cobb-Douglass production function with only one input, labor,

$$Y = AL^{\alpha},$$

where $Y$ is output, $L$ is labor, and $A$ is productivity.

## Discussion of assumptions

- If we take logs, then

  $$\log Y = \log A + \alpha \log L.$$

  If we rearrange slightly, we get something that looks just like a linear regression model,

  $$\underbrace{\log Y_i}_{Y_i} = \underbrace{\mathbb{E}[\log A]}_{\beta_0} + \underbrace{\alpha}_{\beta_1} \underbrace{\log L_i}_{X_i} + \underbrace{(\log A_i - \mathbb{E}[\log A])}_{U_i}.$$

- If this is the model we want to estimate, then $U_i$ is not the error term in the population regression.
- Instead $U_i$ is the difference between the log productivity of firm $i$ and average log productivity.
- It is unlikely that this $U_i$ would be uncorrelated with $\log L_i$.
- More productive firms generally choose to use more inputs, so we should suspect that $U_i$ and $\log L_i$ are positively correlated.

# Discussion of assumptions

SLR.2 Independent observations is a good assumption for data from a simple random sample

- Common situations where it fails in economics are when we have a time series of observations,
- e.g. $\{(X_t, Y_t)\}_{t=1}^{N}$ could be unemployment and GDP of Canada for many different years; and clustering,
- e.g. the data could be students test scores and hours studying and our sample consists of randomly chosen courses or schools—students in the same course would not be independent, but across different courses they might be.
- Still have $\mathbb{E}[\hat{\beta}_1] = \beta_1$ with non-independent observations as long as $\mathbb{E}[Un_i|X_1, ..., X_N] = 0$
- The variance of $\hat{\beta}_1$ will change with non-independent observations

- Independence says that knowing the values of $x_1$ and $y_1$ tells you nothing about the distribution of $x_2$ and $y_2$ (or any other observation).
- When we have cross-sectional data, this assumption usually makes sense.

# Discussion of assumptions

- In economics, we sometimes deal with time-series data, where $(x_1, y_1)$ would be the observation of something at time 1 and $(x_2, y_2)$ is the observation that same thing at time 2. In this case, independence is unlikely to hold.
- Another common situation is panel data, where we observe a sample of individuals over time, so $(x_{it}, y_{it})$ would be what we observe from individual $i$ at time $t$.
- Again, it is unlikely that these observations would be independent over time.
- Later in the course, we will talk about how to deal with non-independent observations.

# Discussion of assumptions

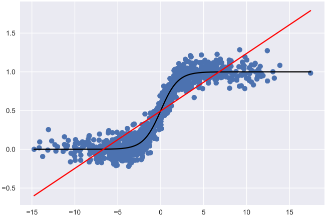
SLR.3 If $\widehat{\mathbb{V}}(X) = 0$, then $\hat{\beta}_1$ involves dividing by 0

- If there is no variation in $X$, then we cannot see how $Y$ is related to $X$

# Discussion of assumptions

SLR.4 To think about mean independence of $U$ from $X$ we should have a model motivating the regression

- If the model we want is just a population regression, then automatically $\mathbb{E}[UX] = 0$, and $\mathbb{E}[U|X] = 0$ if the conditional expectation function is linear; if conditional expectation nonlinear maybe still a useful approximation



## Discussion of assumptions

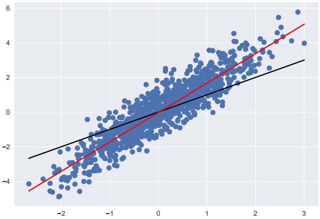
SLR.4 To think about mean independence of $U$ from $X$ we should have a model motivating the regression

- If the model we want is anything else, then maybe $\mathbb{E}[UX] \neq 0$ (and $\mathbb{E}[U|X] \neq 0$), e.g.
    - Demand curve

$$P_i = \beta_0 + \beta_1 Q_i + U_i$$

      $U_i =$ everything that affects price other than quantity. $Q_i$ determined in equilibrium implies $\mathbb{E}[U_i|Q_i] \neq 0$

    - $\mathbb{E}[\hat{\beta}_1] \neq \beta_1$ and $\hat{\beta}_1$ does not tell us what we want



# Discussion of assumptions

- Exogeneity is the most important assumption underlying regression.
- In fact, estimating any economic model using any method will involve some kind of exogeneity assumption.
- By this, we mean that every estimation method requires assuming some error term is either completely independent of some observable ($F_{U|X}(u|x) = F_U(u)$), mean independent of some observable ($\mathbb{E}[U|X] = 0$), or at least uncorrelated with an observable ($\mathbb{E}[UX] = 0$).
- Much of what separates good empirical work in economics from bad is how plausible are the exogeneity assumptions.
- Often, economic theory can help us decide whether or not an exogeneity assumption is plausible. Consider the production function example from earlier,

$$\underbrace{\log Y_i}_{Y_i} = \underbrace{\mathbb{E}[\log A]}_{\beta_0} + \underbrace{\alpha}_{\beta_1} \underbrace{\log L_i}_{X_i} + \underbrace{(\log A_i - \mathbb{E}[\log A])}_{U_i}.$$

## Discussion of assumptions

- To think about whether mean independence of the error term, $\mathbb{E}[\underbrace{(\log A - \mathbb{E}[\log A])} \,|\, \log L] = 0$, makes sense in this model, we should think about how $L$ is determined.
- The firm chooses how much labor to use. Suppose the firm faces output price $P$ and wage $W$.
- If the firm chooses $L$ knowing its productivity, then the firm solves,

$$\max_L PAL^\alpha - WL$$

The first order condition is

$$PA\alpha L^{\alpha-1} - W = 0.$$

If we solve for $A$, we get

$$A = \frac{W}{P\alpha} L^{1-\alpha}$$

$$\log A = \log\left(\frac{W}{P\alpha}\right) + (1-\alpha)\log L$$

## Discussion of assumptions

so

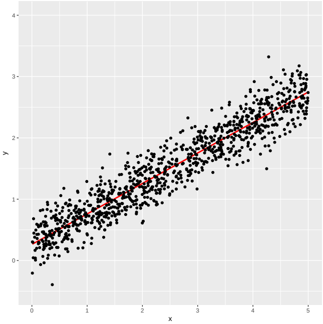
$$\mathbb{E}[\log A | \log L] = (1-\alpha)\log L + \mathbb{E}\left[\log\left(\frac{W}{p\alpha}\right)\middle| \log L\right]$$

- This will not be a constant unless $\alpha = 1$ and $\frac{W}{p}$ is mean independent of $\log L$.
- Both of these are unlikely to hold.
- Unless $\mathbb{E}[\log A | \log L]$ is constant, $\mathbb{E}[\underbrace{(\log A - \mathbb{E}[\log A])}\,|\,\log L] \neq 0$.
- Therefore, exogeneity is not a good assumption in this model, and regression will not give an unbiased estimate of the production function.

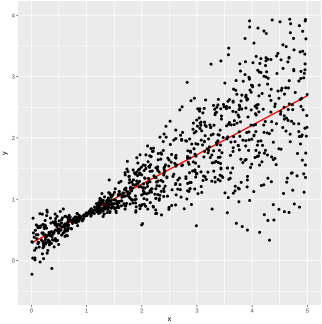
# Discussion of assumptions

SLR.5 Homoskedasticity: variance of $U$ does not depend on $X$

Homoskedastic



Heteroskedastic



- Heteroskedasticity is when $\mathbb{V}(U|X)$ varies with $X$
- If there is heteroskedasticity, the variance of $\hat{\beta}_1$ is different, but we can fix it
- "robust standard errors" / "heteroscedasticity-consistent (HC) standard errors" / "Eicker–Huber–White standard errors"

# Discussion of assumptions

- Homoskedasticity is a strong assumption that is usually not very plausible.
- Therefore, in practice economists almost always calculate heteroscedasticity-robust standard errors.

# Discussion of assumptions

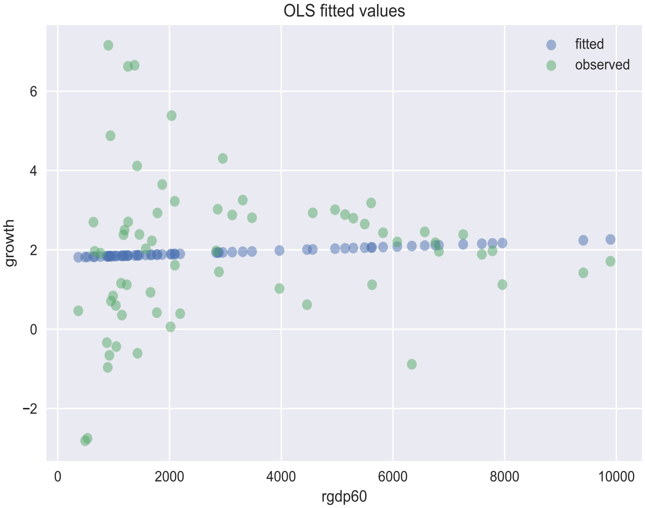
SLR.6 If $U_i|X_i \sim N$, then $\hat{\beta}_1 \sim N$

- What if $U_i$ not normally distributed?
- We will see that $\hat{\beta}_1$ still asymptotically normal

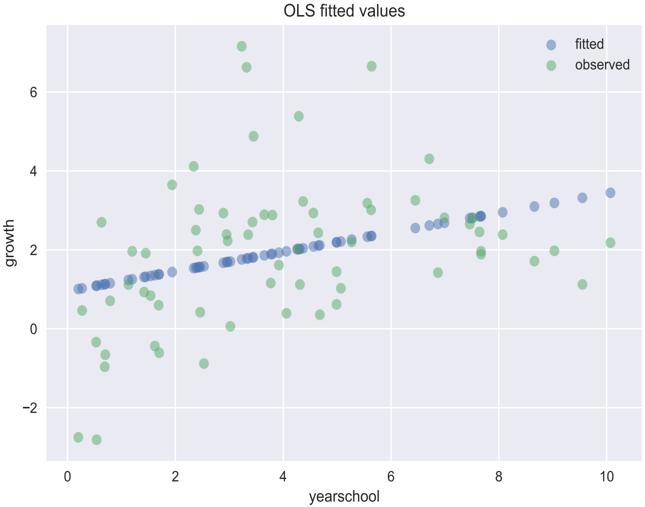
## Example: convergence in growth

- Data on average growth rate from 1960-1995 for 65 countries along with GDP in 1960, average years of schooling in 1960, and other variables
- From http://wps.aw.com/aw_stock_ie_2/50/13016/3332253.cw/index.html, originally used in Beck, Levine, and Loayza (2000)
- Question: has there been in convergence, i.e. did poorer countries in 1960 grow faster and catch-up?

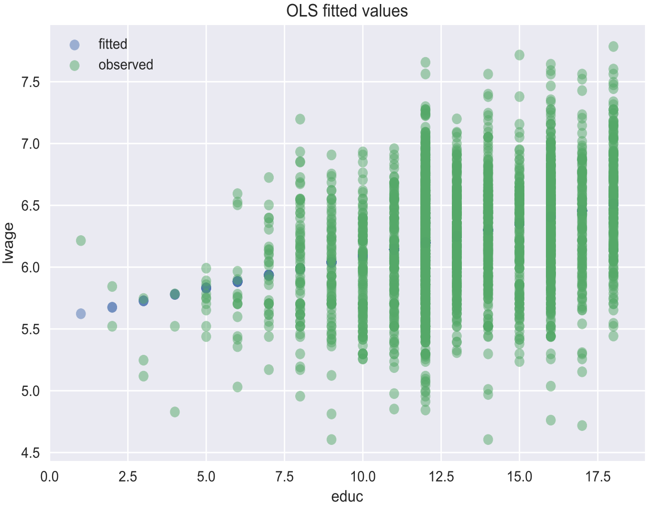
# GDP in 1960 and growth: $\hat{\beta}_0 = 1.7958$ , $\hat{\beta}_1 = 4.735e - 05$.



## Years of schooling in 1960 and growth: $\hat{\beta}_0 = 0.9583$ , $\hat{\beta}_1 = 0.2470$.



# Education and earnings(Card (1993)), : $\hat{\beta}_0 = 5.5709$ , $\hat{\beta}_1 = 0.0521$.



## References

Abbring, Jaap. 2001. "An Introduction to Econometrics: Lecture notes." URL http://jabbring.home.xs4all.nl/courses/b44old/lect210.pdf.

Angrist, J.D. and J.S. Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Angrist, Joshua D and Jörn-Steffen Pischke. 2014. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Baltagi, BH. 2002. *Econometrics*. Springer, New York. URL http://gw2jh3xr2c.search.serialssolutions.com/?sid=sersol&SS_jc=TC0001086635&title=Econometrics.

Beck, T., R. Levine, and N. Loayza. 2000. "Finance and the Sources of Growth." *Journal of financial economics* 58 (1):261–300. URL http://www.sciencedirect.com/science/article/pii/S0304405X00000726.

Bierens, Herman J. 2012. "The Two-Variable Linear Regression Model." URL http://personal.psu.edu/hxb11/LINREG2.PDF.

## References

Card, David. 1993. "Using geographic variation in college proximity to estimate the return to schooling." Tech. rep., National Bureau of Economic Research. URL http://www.nber.org/papers/w4483.

Diez, David M, Christopher D Barr, and Mine Cetinkaya-Rundel. 2012. *OpenIntro Statistics*. OpenIntro. URL http://www.openintro.org/stat/textbook.php.

Stock, J.H. and M.W. Watson. 2009. *Introduction to Econometrics, 2/E*. Addison-Wesley.

Wooldridge, J.M. 2013. *Introductory econometrics: A modern approach*. South-Western.