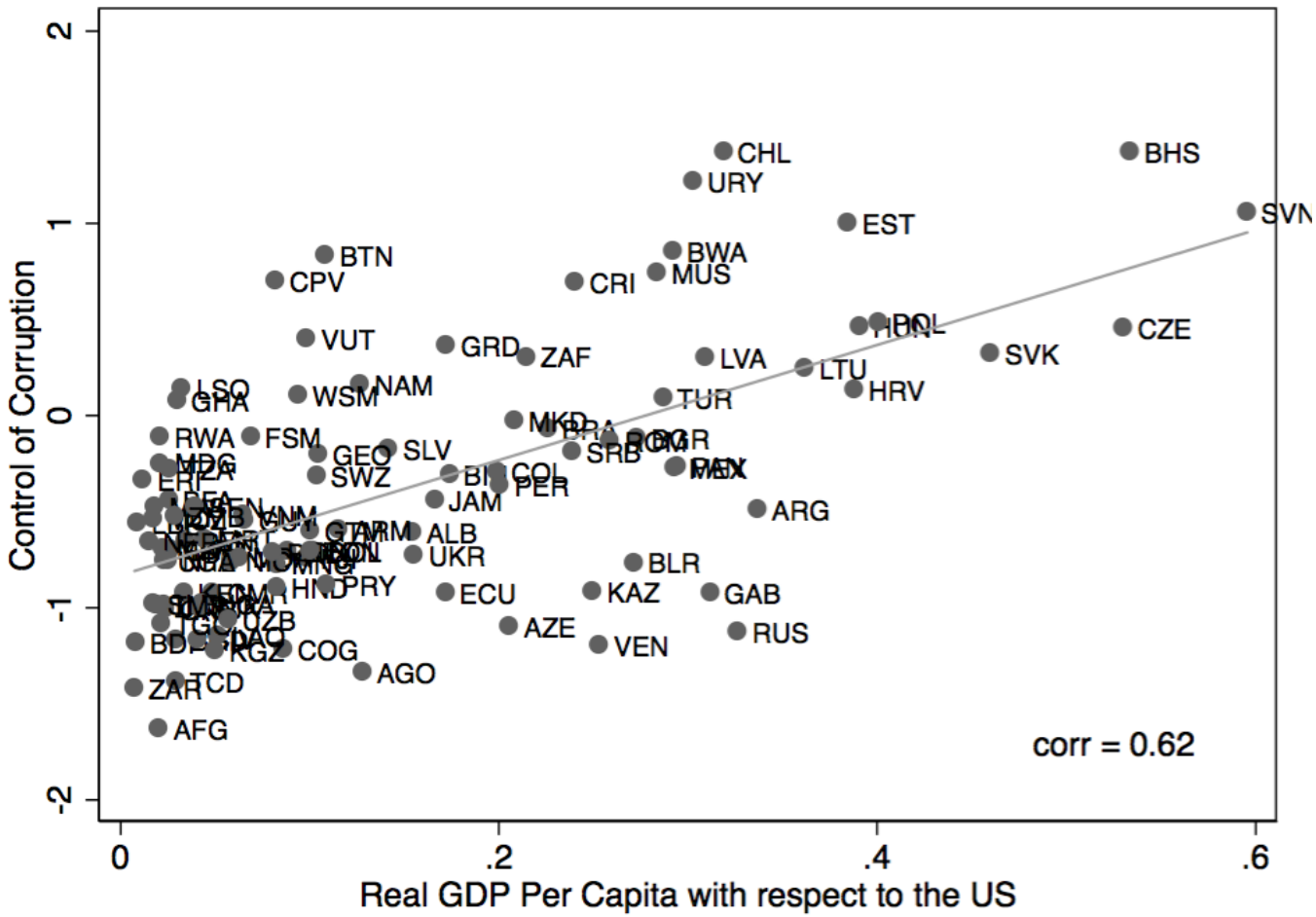
Linear regression with one regressor

Quantitative Methods

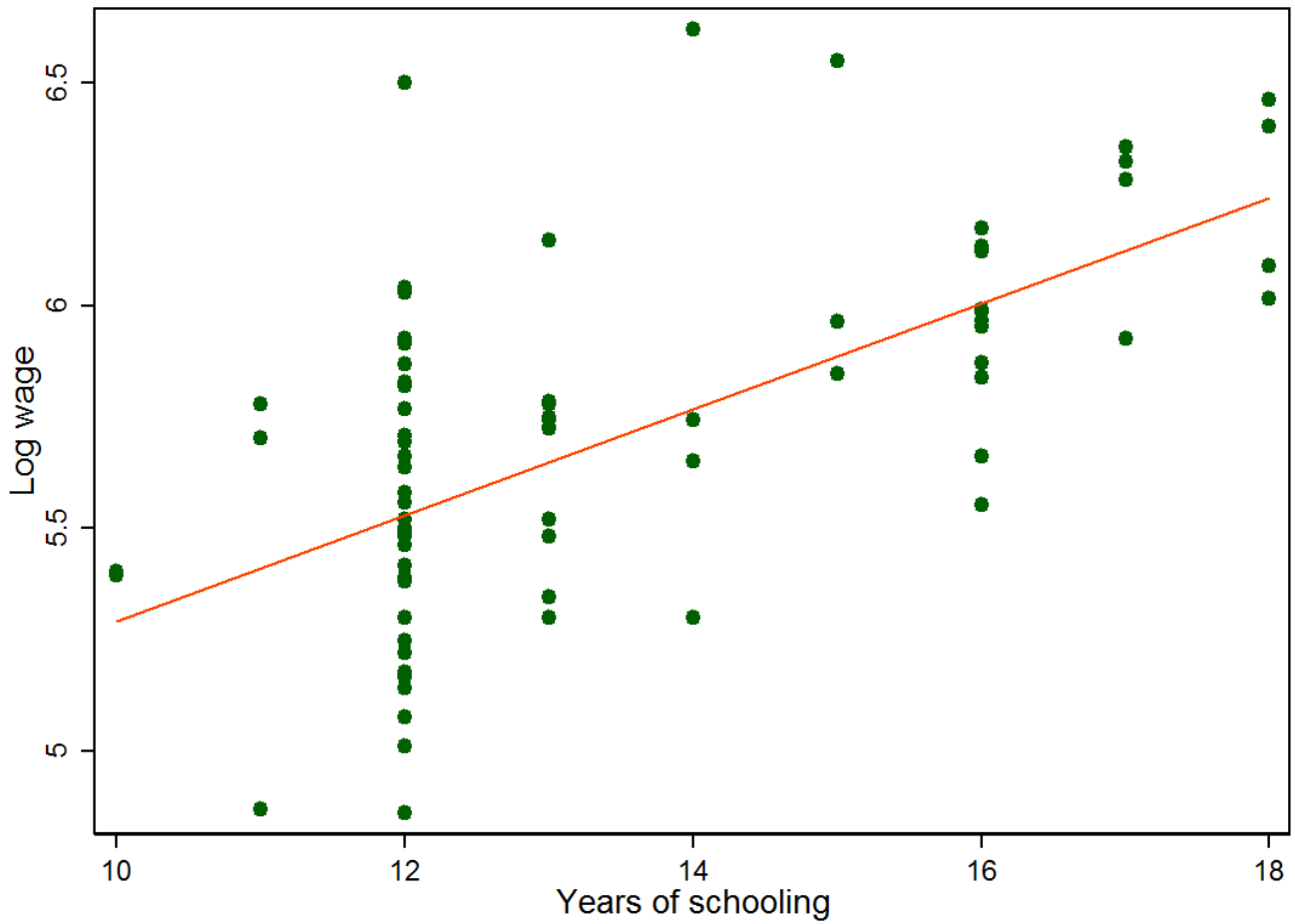
Enrique Moral-Benito

Lesson 3: October 31, 2016

Corruption and the Wealth of Nations



Returns to Schooling



Linear relationships

- In the case of returns to schooling, the covariance between schooling and log wage is 0.50 and the correlation 0.63.
- However, the usefulness of this information per se is rather limited: we only know that they tend to move together.
- A more informative way of representing the fact that two variables y and x are related is to postulate a linear relationship between the two.
- Typically, we think of y as the explained or dependent variable, and of x as the explanatory or independent variable.
- A purely deterministic relationship would be (i.e. the line in the previous scatter plots):

$$y = \beta_0 + \beta_1 x$$

  i.e. a linear function with slope β_1 and intercept β_0.

Linear probabilistic model

- However, looking at the scatter plots we realize that the deterministic relationship is not enough to characterize the observed data points.
- We need a random component capturing the fact that many other factors other than x affect y:

$$y = \beta_0 + \beta_1 x + \epsilon$$

- ϵ is the random error (distance between the straight line and the data points).
- We generally assume that ϵ is unrelated to x and has zero mean:
  - $E(\epsilon|x) = E(\epsilon)$
  - $E(\epsilon) = 0$.
- Then, the deterministic component is also the mean of y: $E(y|x) = \beta_0 + \beta_1 x$
  - If $\beta_1 > 0$, then $E(y|x)$ increases when x increases.
  - If $\beta_1 < 0$, then $E(y|x)$ decreases when x increases.

Ordinary Least Squares (OLS)

- The problem is to estimate the parameters β_0 and β_1, given a random sample $(y_1, x_1, y_2, x_2, ..., y_N, x_N)$ from (y, x).
- So, our aim is to find sample estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$.
- Suppose we fit the straight line $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ to our sample.
- The error of prediction for the ith observation is:

$$\widehat{\epsilon_i} = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

- It makes sense to choose the parameters (i.e, the straight line) that minimizes the errors of prediction.
- As it is not possible to minimize all errors simultaneously, we have to choose a way to aggregate them.
- The ordinary least squares (OLS) estimator chooses the line such that the sum of squared errors is minimal:

$$\sum_{i=1}^{N} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right)^2$$

Ordinary Least Squares (OLS)

- A nice property of OLS is that its solution is analytical.
- FOCs of the minimization problem:

$$\begin{aligned} \sum_{i=1}^{N}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) &= 0 \\ \sum_{i=1}^{N} x_i(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) &= 0 \end{aligned}$$

- Solving for $\widehat{\beta}_0$ and $\widehat{\beta}_1$:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}, \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

  where $\overline{x}$ denotes sample mean, $\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$ the sample covariance and $\sum_{i=1}^{N}(x_i - \overline{x})^2$ the sample variance.
- Note that $\widehat{\beta}_1$ has the same sign as the covariance (and the correlation).

Method of moments

- An alternative interpretation of OLS is based on the method of moments approach.
- Note that $E(\epsilon|x) = E(\epsilon) = 0$ implies that $COV(x, \epsilon) = E(x\epsilon) = 0$. Then:

$$\begin{aligned} E(y - \beta_0 - \beta_1 x) &= 0 \\ E[x(y - \beta_0 - \beta_1 x)] &= 0 \end{aligned}$$

- **Analogy principle**: A natural idea to estimate a quantity in the population is to use the same quantity in the sample (e.g. to estimate $E(y)$ use $\overline{y}$):

$$\begin{aligned} n^{-1} \sum_{i=1}^{N} (y_i - \widehat{\beta}_0 - \widehat{\beta_1} x_i) &= 0 \\ n^{-1} \sum_{i=1}^{N} x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) &= 0 \end{aligned}$$

which coincide with the FOCs for the OLS estimates.

Goodness of fit

- The coefficient of determination measures the explanatory power of the regression model:

$$R^2 = \frac{SSE}{SST} \in [0, 1]$$

  where SST is the Total Sum of Squares and SSE is the explained sum of squares:

$$SST = \sum_{i=1}^{N} (y_i - \overline{y})^2, \qquad SSE = \sum_{i=1}^{N} (\widehat{y}_i - \overline{y})^2$$

- The R^2 is the ratio of the explained variation compared to the total variation.
- It is interpreted as the fraction of the sample variation in y that is explained by x.

Uncertainty in parameter estimates (I)

- Let us consider the case of estimating the sample mean.
- This is equivalent to consider $y_i = \alpha_0 + \epsilon_i$ and estimate $\widehat{\alpha}_0$ by OLS:

$$\overline{y} = \widehat{\alpha}_0 = \frac{1}{N} \sum_{i=1}^{N} y_i$$

- Now, the sample average $\overline{y} = \widehat{\alpha}_0$ is viewed as an estimator instead of as a descriptive statistic.
- Moreover, since each one of the y_i's is a random variable, $\overline{y} = \widehat{\alpha}_0$ is also a **random variable**.
- We typically assume that y_i are drawn independently from the same underlying distribution with variance σ^2 so that:

$$V(\overline{y}) = V(\widehat{\alpha}_0) = \frac{\sigma^2}{N}.$$

Uncertainty in parameter estimates (II)

- Let us check these results in STATA.
- We first compute $\widehat{\alpha}_0$ from a randomly generated sample with 20 observations:
- `#delimit ;`
- `clear all; set obs 20; gen y = 3 + rnormal(); sum y; reg y;`
- Now, to understand what $\widehat{\alpha}_0$ being random actually means, we can perform the following experiment:
  - Suppose we draw another sample from the same population and estimate $\widehat{\alpha}_0$:
  - `clear all ; set obs 20 ; gen y = 3 + rnormal() ; sum y ; reg y ;`
  - The resulting sample mean will be different.
  - In this sense, the sample mean estimator is a random variable.
- Note that we have just conducted a **Monte Carlo** simulation!
- The frequentist approach to statistics studies the property of estimators under this thought experiment. That is, under repeated sampling.

Monte Carlo and bootstrap

- A **Monte Carlo** simulation is based on simulating various samples from the population, and computing the estimator on each of those samples.
- Formally:
  - Simulate a large number S of random samples $y_1^{(s)}, ..., y_N^{(s)}$ from the population.
  - For each s, compute $\overline{y}^{(s)} = \widehat{\alpha}_0^{(s)} = 1/N \sum y_i^{(s)}$.
- The idea of the procedure is that, when S is very large, the sequence $\widehat{\beta}_0^{(1)}, ..., \widehat{\beta}_0^{(S)}$ will replicate the distribution of the random variable $\widehat{\alpha}_0$.

- In many applications, however, we do not observe the full population.
- The idea of the **bootstrap** is to draw (with replacement) random observations from the sample that we observe.

Finite samples: Unbiasedness and efficiency

- We generally want the estimator to be close to the population value.
- We ask that an estimator be good on average, that is across repeated samples.
- There are two main criteria in finite samples:
- **Unbiasedness**:
  - Is the mean of the estimator equal to the population value?
  - The sample mean is unbiased: $E(\widehat{\alpha}_0) = \alpha_0$.
  - We can check that the mean of the sample means across an infinite number of Monte Carlo simulations is indeed equal to the population mean.
- **Efficiency**:
  - Is the variance of the estimator smaller than for alternative (unbiased) estimators?
  - This criterion requires that the distribution of the estimator be tightly concentrated around the true value.
  - The sample mean is efficient: $V(\widehat{\alpha}_0) = \frac{\sigma^2}{N}$.

Large samples: Consistency

- A different criterion for good estimation is that the estimator behave well when the sample size N increases.
- Intuitively, when N increases, the sample gets closer to the population, so the estimator should converge to the true value.
- **Consistency**:
  - If $\widehat{\alpha}_0$ is an unbiased estimator of α_0 and $V(\widehat{\alpha}_0) \to 0$ as $N \to \infty$, then $\widehat{\alpha}_0$ is consistent.
  - We can easily check in STATA that the sample mean is consistent for the population mean.
  - This is an application of the law of large numbers.
- **Central Limit Theorem**:
  - Let $(y_1, y_2, ..., y_N)$ be a random sample with mean μ and variance σ^2. Then:

$$Z = \frac{\overline{y} - \mu}{\sigma/\sqrt{N}}$$

  - Has an asymptotic standard normal distribution.

Comparing estimators: an example

- Suppose we have a random sample $x_1, x_2, ..., x_{50}$ from the population.
- We can estimate the expected value of X, namely, μ with the following alternative estimators:

$$\begin{aligned}
\widehat{\mu}_1 &= x_1 \\
\widehat{\mu}_2 &= \frac{x_1 + x_2}{2} \\
\widehat{\mu}_3 &= \frac{x_1 + x_2}{3} \\
\widehat{\mu}_4 &= \frac{\sum_{j=1}^{50} x_j}{50}
\end{aligned}$$

- Which is unbiased?
- Which is the most efficient?
- Which is consistent?

Maximum likelihood

- Maximum-likelihood estimation (MLE) is an alternative method of estimating the parameters of a model given data and a distributional assumption.
- For instance, we can assume that y is normal with mean μ and variance σ^2.

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

  where $\theta = (\mu, \sigma^2)$ are the parameters of the model.
- We define the log-likelihood $L(y; \theta) = \log f(y; \theta)$, which is a random variable.
- The ML estimator maximizes expectation of $L(y; \theta)$:

$$\widehat{\theta}_{ML} = \underset{c}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} L(y_i; c)$$

- Intuitively, ML finds particular parameter values that make the observed data the most probable given the model.
- Under mild conditions, ML estimates are consistent and asymptotically normal.
- Moreover, ML is the most asymptotically efficient estimator when the model is correctly specified.

Confidence intervals (I)

- A point estimate from a particular sample does not provide enough information for informing policy discussions.
- Suppose we have a random sample $y_1, y_2, ..., y_N$ from a population with a $N(\mu, \sigma^2)$ distribution.
- In this case, the sample mean is distributed according to $N(\mu, \sigma^2/N)$.
- Using some results on the normal distribution, this implies that:

$$P\left(-1.96 < \frac{\overline{y} - \mu}{\sigma/\sqrt{N}} < 1.96\right) = 0.95$$

- Therefore, $\overline{y}$ will belong to the interval:

$$\left(\overline{y} - 1.96\sigma/\sqrt{N}, \overline{y} + 1.96\sigma/\sqrt{N}\right)$$

  approximately 95% of the times (in repeated samples).
- This interval is a very useful measure of uncertainty of the sample mean, and is referred to as a 95% confidence interval.

Confidence intervals (II)

- The probabilistic interpretation of a 95% confidence interval is that:
  - For 95% of all random samples, the constructed interval will contain μ.
  - We can check this in the computer by means of a Monte Carlo simulation.
- Confidence intervals are easy to compute from regression outputs.
- The Central Limit Theorem is very useful to construct confidence intervals.
- A rule of thumb for an approximate 95% confidence interval is:

$$[\overline{y} \pm 1.96 \cdot se(\overline{y})]$$

Hypothesis Testing

- Sometimes the question we are interested in has a definite yes or no answer.
- For instance, we might want to test a null hypothesis such as:

$$H_0 : \mu = \mu_0$$

- In this case, we can use the statistic $T = \frac{\overline{y} - \mu_0}{se(\overline{y})}$, which is asymptotically normal under H_0.
- The p-value is the probability of obtaining a result equal to or "more extreme" than what was actually observed, assuming that H_0 is true.
  - If we obtain $T = 2.5$, then $p = 0.006$.
  - If we obtain $T = 0.5$, then $p = 0.308$.
- Small p-values are evidence against H_0, since they indicate that the outcome of the data occurs with small probability if H_0 is true.
- Confidence intervals can be used to test hypotheses about parameters:
  - If μ_0 is not in the 95% confidence interval, then H_0 is rejected at the 5% level.
  - If μ_0 lies in this interval, then we fail to reject H_0.

Statistical versus economic significance, and nonlinearities

- A statistically significant effect may be economically insignificant and viceversa.
- Statistical significance can be driven from a large estimate or a small standard error (which may result from a large sample size).
- Economic significance can be assessed by comparing the estimate with the variable of interest.
- The magnitude of the coefficients depends on the unit of measurement of the dependent and independent variables.
- Nonlinearities can be incorporated to the linear regression model:
  - level-level: $\Delta y = \beta_1 \Delta x$
  - level-log: $\Delta y = (\beta_1/100)\%\Delta x$
  - log-level: $\%\Delta y = (100\beta_1)\Delta x$
  - log-log: $\%\Delta y = \beta_1 \%\Delta x$
- Note that the model must be linear in parameters (in this course).

Correlation versus causation

- An important remark, which you will see in the **econometrics** course, is that regression results are evidence of an association between some variables.
- In general, they must not be interpreted as x causing y.

- Example 1:
  - Since the 1950s, both the atmospheric CO2 level and obesity levels have increased sharply.
  - Hence, atmospheric CO2 causes obesity.
  - Explanation: richer populations tend to eat more food and consume more energy.

- Example 2:
  - Sleeping with one's shoes on is strongly correlated with waking up with a headache.
  - Therefore, sleeping with one's shoes on causes headache.
  - Explanation: both are caused by a third factor, in this case going to bed drunk.

- The relevance of this concern depends on the objective of our analysis.
- See http://nadaesgratis.es/libertad-gonzalez/para-que-sirve-la-econometria