Section 2 Simple Regression

What regression does

- Relationship between variables
  - Often in economics we believe that there is a (perhaps causal) relationship between two variables.
  - Usually more than two, but that's deferred to another day.
  - We call this the ***economic model***.
  - **Example:** Grade in Econ 201 vs. number of dorm-mates taking Econ 201
- Functional form
  - Is the relationship linear?
    - $y = \beta_0 + \beta_1 x$
    - x is called the "regressor"
    - The linear form is a natural first assumption, unless theory rejects it.
    - β_1 is the slope, which determines whether relationship between x and y is positive or negative.
      - $\frac{dy}{dx} = \beta_1$
    - β_0 is the intercept or constant term, which determines where the linear relationship intersects the y axis.
  - Is it plausible that this is an exact, "deterministic" relationship?
    - No. Data (almost) never fit exactly along line.
    - Why?
      - Measurement error (incorrect definition or mismeasurement)
      - Other variables that affect y
      - Relationship is not purely linear
      - Relationship may be different for different observations
  - So the economic model must be modeled as determining the *expected value* of y
    - $E(y \mid x) = \beta_0 + \beta_1 x$: The *conditional mean* of y given x is $\beta_0 + \beta_1 x$
      - Note that this says nothing about other aspects of the distribution (other than the expected value)
      - How does a change in x affect the variance of y? (We assume that it does not.)
      - How does a change in x affect the median, or the 75th percentile, or any other aspect of the distribution of y? (If y is assumed to be normal, then everything about the distribution depends only on mean and variance.)

- Other regression techniques (in particular, quantile regression) allow us to examine the impact of *x* on aspects of the distribution of *y* other than the mean.

- Adding an (unexplained) error or disturbance term for a "stochastic" relationship gives us the actual value of *y*: $y = \beta_0 + \beta_1 x + u$
- Error term *u* captures all of the above problems.
  - Error term is considered to be a random variable and is not observed directly.
  - We must assume that $E(u \mid x) = E(u) = 0$
    - This means that the mean of the error term is zero, *regardless of the value of x*.
    - It really nests two assumptions:
      - That the average error term is zero
      - That the error term is independent of *x*
  - Variance of *u* is σ^2, which is the *conditional variance* of *y* given *x*, the variance of the conditional distribution of *y* given *x*, or $\text{var}(y \mid x) = \text{var}(u \mid x) = \sigma^2$.
  - This is the simplest, but often not valid, assumption: that the conditional variance is the same for all observations in our sample (*homoskedasticity*)
- $\beta_1 = \frac{dE(y \mid x)}{dx}$, which means that the expected value of *y* increases by β_1 units when *x* increases by one unit

- Does it matter which variable is on the left-hand side?
  - At one level, no:
    - $x = \frac{1}{\beta_1}(y - \beta_0 - u)$, so
    - $x = \gamma_0 + \gamma_1 y + v$, where $\gamma_0 \equiv -\frac{\beta_0}{\beta_1}, \quad \gamma_1 \equiv \frac{1}{\beta_1}, \quad v \equiv -\frac{1}{\beta_1} u.$
  - For purposes of most estimators, yes:
    - We shall see that a critically important assumption is that the error term is independent of the "regressors" or *exogenous* variables.
    - Are the errors shocks to *y* for given *x* or shocks to *x* for given *y*?
      - It might not seem like there is much difference, but the assumption is crucial to valid estimation.
      - This is the crucial role of $E(u \mid x) = 0$ because if *y* affects *x*, then *u* and *x* are correlated and $E(u \mid x) \neq 0$

        - It cannot be the case that $E(u|x)=0$ **and** $E(u|y)=0$.
    - Exogeneity: *x* is exogenous with respect to *y* if shocks to *y* do not affect *x*, i.e., *y* does not cause *x*.
- Where do the data come from? Sample and "population"
    - We observe a sample of observations on *y* and *x*.
    - Depending on context these samples may be
        - Drawn from a larger population, such as census data or surveys
        - Generated by a specific "data-generating process" (DGP) as in time-series observations
    - We usually would like to assume that the observations in our sample are statistically independent, or at least uncorrelated: $\text{cov}(y_i, y_j)=0, \forall i \neq j$.
    - IID samples have independently and identically distributed observations.
- Goals of regression
    - True regression line: actual relationship in population or DGP
        - True β and $f(u|x)$
        - Sample of observations comes from drawing random realizations of *u* from $f(u|x)$ and plotting points appropriately above and below the true regression line.
    - We want to find an estimated regression line that comes as close to the true regression line as possible, based on the observed sample of *y* and *x* pairs:
        - Estimate values of parameters β_0 and β_1
        - Estimate properties of probability distribution of error term *u*
        - Make inferences about the above estimates
        - Use the estimates to make conditional forecasts of *y*
        - Determine the statistical reliability of these forecasts

Summarizing assumptions of simple regression model

- **Assumption #0**: (Implicit and unstated) The model as specified applies to all units in the population and therefore all units in the sample.
    - All units in the population under consideration have the same form of the relationship, the same coefficients, and error terms with the same properties.
    - If the United States and Mali are in the population, do they really have the same parameters?
    - This assumption underlies everything we do in econometrics, and thus it must always be considered very carefully in choosing a specification and a sample, and in deciding for what population the results carry implications.
- SLR1: $y=\beta_0+\beta_1 x+u$
- SLR2: Random sampling

    - We have random sample of (y_i, x_i) for $i = 1, 2, \ldots, n$ that comes from population with SLR1.
- SLR3: x takes on at least two distinct values
    - We need this because we can't estimate the slope if the observations are all perfectly aligned vertically.
- SLR4: $E(u \mid x) = 0$
    - As noted above, this incorporates both the assumption that u is independent of x and the assumption that u has zero overall mean.
- SLR5: Homoskedasticity, or $\text{var}(u \mid x) = \sigma^2$
- We sometimes also assume that the error term follows a normal distribution
- **Example:** Assess the validity of these assumptions for 201 dorm-mate model

Introduction to Stata

- Stata works on a dataset (.dta file)
- Stata commands:
    - Enter at prompt
    - Choose from menu/windows
    - Enter into a do file for batch execution
- The Stata screen
    - Results window
    - Command window
    - Variables window
    - Review window
    - Properties window
- Log files
    - Set one up so students can see it later
- Opening a data set
    - Show data editor/browser
- Commands to do statistical analysis
    - summarize
    - reg
- Graphics commands
    - Use menus to get see all options without remembering how to type
- Sample analysis: Reed Econ 201 grades
    - Dependent variable gpoints
        - **Start log file**
        - Show summary statistics
        - **Summarize gpoints hsgpa**
            - Note missing data for hsgpa: Is this a problem?

    - **Twoway (scatter gpoints hsgpa)**
        - Point out discrete distribution: Is this a problem?
        - Point out outlier in hspga: Is this a problem?
- Regression on single variable: **regress gpoints hsgpa**
    - Interpreting coefficients (note that intercept is automatically included: noint option)
    - Point out standard error, t statistic, p value, confident limits
    - Note missing observations: **misstable patterns gpoints hsgpa**
    - Show **outreg2 using graderegs, word**
    - **estimates store** is available, but vastly inferior
- Alternative: **regress gpoints irdr**
    - Show how outreg adds columns
        - **outreg2 using graderegs, word**
        - Note parallel columns
        - Note difference in number of observations
    - Calculate predicted values with predict
        - **predict gpahat if e(sample)**
        - **predict uhat if e(sample) , resid**
        - Graph actual and predicted vs. irdr
            - **twoway (scatter gpahat irdr)**
            - **twoway (scatter gpahat irdr) (scatter gpoints irdr)**
            - **twoway (scatter uhat irdr)**
    - Display hypothetical predicted values with margins
        - **margins , at(irdr=(5 4 3 2))**
- Transformation: **generate satc100 = satv100 + satm100**
    - **Regress gpoints satc100**
    - Compare *n* to hsgpa regression
- Regression on dummy variable
    - **Regress gpoints female**
    - Interpretation of coefficients
    - Category mean predictions:
        - **margins female**
- Multiple regression demonstration
    - **Reg gpoints irdr satv100 satm100 female**
    - Show outreg with multiple variables
        - **outreg2 using graderegs , word**
    - Add **taking** to regression and interpret
    - Use margins to isolate predictions of hypothetical individual variables with others at means
        - **margins , at(irdr = (5 4 3 2)) atmeans**
        - **marginsplot**
- **Don't forget to save log file and outreg file on flash drive**

Strategies for obtaining regression estimators

- What is an *estimator*?
  - A rule (formula) for calculating an *estimate* of a parameter (β_0, β_1, or σ^2) based on the sample values y, x
  - Estimators are often denoted by ^ over the variable being estimated: An estimator of β_1 might be denoted $\hat{\beta}_1$
- How might we estimate the β coefficients of the simple regression model?
  - Three strategies:
    - Method of least-squares
    - Method of maximum likelihood
    - Method of moments
  - All three strategies with the SR assumptions lead to the same estimator rule: the *ordinary least-squares* regression estimator: $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$
- The section builds on Section C-4 of Wooldridge's statistics review, pages 724–726.
- **Method of least squares**
  - Estimation strategy: Make sum of squared y-deviations ("residuals") of observed values from the estimated regression line as small as possible.
  - Given coefficient estimates $\hat{\beta}_0, \hat{\beta}_1$, residuals are defined as $\hat{u}_i \equiv y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
    - Or $\hat{u}_i = y_i - \hat{y}_i$, with $\hat{y}_i \equiv \hat{\beta}_0 + \hat{\beta}_1 x_i$
  - Why not minimize the sum of the residuals?
    - We don't want sum of residuals to be large negative number: Minimize sum of residuals by having all residuals infinitely negative.
    - Many alternative lines that make sum of residuals zero (which is desirable) because positives and negatives cancel out.
  - Why use square rather than absolute value to deal with cancellation of positives and negatives?
    - Square function is continuously differentiable; absolute value function is not.
      - Least-squares estimation is much easier than least-absolute-deviation estimation.
    - Prominence of Gaussian (normal) distribution in nature and statistical theory focuses us on variance, which is expectation of square.
    - Least-absolute-deviation estimation is occasionally done (special case of quantile regression), but not common.
    - Least-absolute-deviation regression gives less importance to large outliers than least-squares because squaring gives large emphasis to residuals with large absolute value. Tends to draw the regression line toward these points to eliminate large squared residuals.

- Least-squares criterion function: $SST \equiv \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$
  - Least-squares estimators is the solution to $\min_{\hat{\beta}_0, \hat{\beta}_1} SST$. Since SST is a continuously differentiable function of the estimated parameters, we can differentiate and set the partial derivatives equal to zero to get the **least-squares normal equations:**
    - $$\frac{\partial SST}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)(-x_i) = 0,$$
    - $$-\sum_{i=1}^{n} y_i x_i + \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0.$$
    $$\frac{\partial SST}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} -2\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$
    - $$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$
    $$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} = 0$$
    $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$
  - Note that the $\hat{\beta}_0$ condition assures that the regression line passes through the point $(\bar{x}, \bar{y})$.
  - Substituting the second condition into the first divided by *N*:
    $$-\sum y_i x_i + \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) n\bar{x} + \hat{\beta}_1 \sum x_i^2 = 0$$
    $$-\left(\sum y_i x_i - n\overline{yx} \right) + \hat{\beta}_1 \left(\sum x_i^2 - n\bar{x}^2 \right) = 0$$
    $$\hat{\beta}_1 = \frac{\sum y_i x_i - n\overline{yx}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}.$$
  - The $\hat{\beta}_1$ estimator is the sample covariance of *x* and *y* divided by the sample variance of *x*.
  - What happens if *x* is constant across all observations in our sample?
    - Denominator is zero and we can't calculate $\hat{\beta}_1$.
    - This is our first encounter with the problem of collinearity: if *x* is a constant then *x* is a linear combination of the "other regressor"—the constant one that is multiplied by $\hat{\beta}_1$.
    - Assumption SLR3 rules out this possibility.
    - Collinearity (or multicollinearity) will be more of a problem in multiple regression. If it is extreme (or perfect), it means that we can't calculate the slope estimates.

- The above equations are the "ordinary least-squares" (OLS) coefficient estimators.
- The two partial derivative equations set to zero are called the "**least-squares normal equations**."
  - We shall see that these same equations result from the other two methods we consider, and therefore that the other methods also lead to the same estimators.

- **Method of maximum likelihood**
  - Consider the joint probability density function of y_i and x_i, $f_i(y_i, x_i \mid \beta_0, \beta_1)$. The function is written as conditional on the coefficients β to make explicit that the joint distribution of y and x is affected by the parameters.
    - This function measures the probability density of any particular combination of y and x values, which can be loosely thought of as how "likely" that outcome is, given the parameter values.
    - For a given set of parameters, some observations of y and x are less likely than others. For example, if $\beta_0 = 0$ and $\beta_1 < 0$, then it is less likely that we would see observations where $y > 0$ than observations with $y < 0$ when $x > 0$,.
  - The idea of maximum-likelihood estimation is to choose as our maximum-likelihood estimator the set of parameters that makes the likelihood of observing the sample that we actually have as high as possible.
  - The *likelihood function* is just the joint density function turned on its head: $L_i\left(\hat{\beta}_0, \hat{\beta}_1 \mid x_i, y_i\right) \equiv f_i\left(x_i, y_i \mid \beta_0, \beta_1\right)$.
    - We add the hats because we will want to calculate estimators for the β coefficients based on maximizing this equation
  - If the observations are independent random draws from identical probability distributions (they are IID), then the overall sample density (likelihood) function is the product of the density (likelihood) function of the individual observations:
    - $$f\left(x_1, y_1, x_2, y_2, \ldots, x_n, y_n \mid \beta_0, \beta_1\right) = \prod_{i=1}^{n} f_i\left(x_i, y_i \mid \beta_0, \beta_1\right)$$
    - $$L\left(\hat{\beta}_0, \hat{\beta}_1 \mid x_1, y_1, x_2, y_2, \ldots, x_n, y_n\right) = \prod_{i=1}^{n} L_i\left(\hat{\beta}_0, \hat{\beta}_1 \mid x_i, y_i\right).$$
  - If the conditional probability distribution of u conditional on x is Gaussian (normal) with mean zero and variance σ^2:
    - $$f_i\left(x_i, y_i \mid \beta_0, \beta_1\right) = L_i\left(\hat{\beta}_0, \hat{\beta}_1 \mid x_i, y_i\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(\frac{-\frac{1}{2}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{\sigma^2}\right)}$$
    - Because of the exponential function, Gaussian likelihood functions are usually manipulated in logs.

- Note that because the log function is monotonic, maximizing the log-likelihood function is equivalent to maximizing the likelihood function itself.

- For an individual observation: $\ln L_i = -\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$
- Aggregating over the sample:

$$\ln \prod_{i=1}^{n} L_i\left(\hat{\beta}_0, \hat{\beta}_1 \mid x_i, y_i\right) = \sum_{i=1}^{n} \ln L_i\left(\hat{\beta}_0, \hat{\beta}_1 \mid x_i, y_i\right)$$
$$= \sum_{i=1}^{n}\left[-\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2\right]$$
$$= -\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2.$$

  - The only part of this expression that depends on $\hat{\beta}$ or on the sample is the final summation. Because of the negative sign, maximizing the likelihood function (with respect to β) is equivalent to minimizing the summation.
    - But this summation is just the sum of squared residuals that we minimized in OLS.
  - Thus, OLS is MLE if the distribution of *u* conditional on *x* is Gaussian with mean zero and constant variance σ^2, and if the observations are IID.

- **Method of moments**
  - Another general strategy for obtaining estimators is to set estimates of selected population moments equal to their sample counterparts. This is called the method of moments.
  - In order to employ the method of moments, we have to make some specific assumptions about the population/DGP moments, both of which are implied by SLR4.
    - Assume $E\left(u_i\right) = 0, \forall i$. This means that the population/DGP mean of the error term is zero.
      - Corresponding to this assumption about the population mean of *e* is the sample mean condition $\frac{1}{n}\sum \hat{u}_i = 0$. Thus we set the sample mean to the value we have assumed for the population mean.
    - Assume $\text{cov}\left(x,u\right) = 0$, which is equivalent to $E\left[\left(x_i - E(x)\right)u_i\right] = 0.$
      - Corresponding to this assumption about the population covariance between the regressor and the error term is the sample covariance condition: $\frac{1}{n}\sum\left(x_i - \bar{x}\right)\hat{u}_i = 0.$ Again, we set the sample moment to the zero value that we have assumed for the population moment.

- Plugging the expression for the residual into the sample moment expressions above:
  - $\frac{1}{n}\sum\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0,$
    $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$
  - This is the same as the intercept estimate equation for the least-squares estimator above.
  - $\frac{1}{n}\sum(x_i - \bar{x})\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0,$
    $\sum(x_i - \bar{x})\left(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i\right) = 0,$
    $\sum(x_i - \bar{x})(y_i - \bar{y}) - \sum\hat{\beta}_1(x_i - \bar{x})(x_i - \bar{x}) = 0,$
    $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}.$
  - This is exactly the same equation as for the OLS estimator.
- Thus, if we assume that $E(u_i) = 0, \forall i$ and $\text{cov}(x,u) = 0$ in the population, then the OLS estimator can be derived by the method of moments as well.
- **By construction:**
  - The mean of the residuals is zero
  - The sample correlation of the residuals with *x* is zero
  - These two conditions **are** the OLS normal equations and the OLS estimators have these properties regardless of what assumptions are or are not true in the actual data
  - **Only** if the corresponding assumptions about population moments (SLR4) are true in the actual population will we get reliable estimators from assuming them in the sample
- (Note that both of these moment conditions follow from the extended assumption SLR4 that $E(u \mid x) = 0$.)

- Evaluating alternative estimators (not important for comparison here since all three are same, but are they any good?)
  - Desirable criteria
    - Unbiasedness: estimator is on average equal to the true value
      - $E\left(\hat{\beta}\right) = \beta$
    - Small variance: estimator is usually close to its expected value
      - $\text{var}\left(\hat{\beta}\right) = E\left[\left(\hat{\beta} - E\hat{\beta}\right)^2\right]$
    - Small RMSE can balance variance with bias:

- $RMSE = \sqrt{MSE}$
- $MSE \equiv E\left[\left(\hat{\beta} - \beta\right)^2\right]$
    - We will talk about BLUE estimators as minimum variance within the class of unbiased estimators.

Sums of squares and goodness of fit

- We know that $y_i = \hat{y}_i + \hat{u}_i$ and that, by construction, $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u} = 0$ and the sample covariance between $\hat{y}$ and $\hat{u}$, $\frac{1}{n-1}\sum_{i=1}^{n}\hat{u}_i\left(x_i - \bar{x}\right) = 0.$
- These conditions imply that we can decompose SST (total sum of squares) into two independent parts:
    - SST = SSE + SSR, with
        - $SST = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$,
        - $SSE = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$, (=explained sum of squares) and
        - $SSR = \sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2 = \sum_{i=1}^{n}\hat{u}^2$. (=residual sum of squares)
    - This works because of the two moment conditions, as Wooldridge proves on page 34.
    - Note ambiguity of notation:
        - Some authors use SSR to be the "regression sum of squares" and SSE to be the "error sum of squares"
        - You have to look carefully when reading textbooks or Stata documentation to see which convention the author uses
        - Stata uses "Model SS" (mss) and "Residual SS" (rss)
- One fundamental question is how well our model fits the data
    - Small (absolute) residuals would indicate a good fit
    - $R^2 \equiv \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$ is a standard measure of goodness of fit
        - If residuals are all zero, then the fit is perfect, SSR = 0, and SSE = SST, so $R^2 = 1$
        - If the model explains nothing, then our best estimate for each observation is $\hat{y}_i = \bar{y}$, and noting the definitions, SSE = 0, so SSR = SST and $R^2 = 0$.
    - Thus, R^2 gives a scale-independent measure of goodness of fit that is always between 0 and 1

Units of measure and scaling

- There are lots of units we might choose for economic variables
  - GDP can be in billions or trillions
  - Eggs can be in units, dozens, or pounds
  - Prices can be $/unit where unit can be anything, or an index number that can be centered at 1 or 100
- What difference does this make for our OLS estimators?
  - If we **add or subtract** a constant from either *x* or *y*, all that is affected is the intercept term $\hat{\beta}_0$. Since we are not usually very interested in the value of the intercept, this is usually meaningless.
  - If we **multiply *x*** by a constant, the slope estimate $\hat{\beta}_1$ will be *divided* by the same constant (as will its standard error, leaving the *t* statistic unchanged). The estimated intercept is unchanged, as are the residuals, $\hat{\sigma}^2$, and R^2.
  - If we **multiply *y*** by a constant, the slope and intercept estimates will both be multiplied by the same constant (as will their standard errors, leaving the *t* statistics unchanged) and $\hat{\sigma}^2$.
  - None of these transformations has any effect on R^2.
- Simple nonlinearities:
  - What is crucial about the linear regression model is that it is linear in the coefficients
  - The variables do not need to be linear: *x* could for example be ln*z*, where *z* is a variable whose effect on *y* is proportional. Similarly, *y* could be the log of another variable.
  - Since $d \ln y = \frac{dy}{y}$, the change in the natural log of *y* is (for very small changes) equal to the **proportional** change in *y*
    - That means that if $\ln y_i = \beta_0 + \beta_1 \ln x_i + u_i$, then β_1 is the **elasticity** of *y* with respect to *x*
    - We use the log-log functional form a lot in econometrics because we are often interested in estimating elasticities
    - Wooldridge's Table 2.3 summarizes the four cases of levels and logs of *y* and *x* and how we would interpret the coefficients

Sampling distribution of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables: they are functions of the random variables *y* and *u*.
  - We can think of the probability distribution of $\hat{\beta}$ as occurring over repeated random samples from the underlying population or DGP.

- In many (most) cases, we cannot derive the distribution of an estimator theoretically, but must rely on Monte Carlo simulation to estimate it. (Discussed below)
    - Because OLS estimator (under our assumptions) is linear, we *can* derive its distribution
- We can write the OLS slope estimator as

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

$$= \frac{\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + u_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i + u_i - (\beta_0 + \beta_1 \bar{x}))(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(\beta_1 (x_i - \bar{x}) + u_i)(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^{n} u_i (x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  The third step uses the property $\bar{y} = \beta_0 + \beta_1 \bar{x}$, since the expected value of *e* is zero.
- We are interested in the moments of the distribution of the OLS estimator, its mean and variance
    - We take these moments **conditional on *x***
    - Because of the conditionality, we can treat the *x* values as known constants
    - Because all the *x* variables are treated as known, they can come outside when we take conditional expectations, so

$$E(\hat{\beta}_1 \mid x) = \beta_1 + E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \middle| x\right] = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(u_i \mid x)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1 .$$

        - We use SLR4 in the last step to know that $E(u_i \mid x) = 0$
    - This shows that the OLS estimator is **unbiased** as long as SLR4 holds
    - What about the **variance** of $\hat{\beta}_1$?

$$\text{var}\left(\hat{\beta}_1\right) = E\left[\left(\hat{\beta}_1 - \beta_1\right)^2 \middle| x\right]$$

- $$= E\left[\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})E(u_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]^2 \middle| x\right]$$

$$= \ldots$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Wooldridge Theorems 2.1 and 2.2 provide formulas for mean and variance of $\hat{\beta}_0$ and we also are interested in the covariance between the coefficients:
  - $E\left(\hat{\beta}_0 \mid x\right) = \beta_0$, so the estimate of the constant is unbiased
  - $\text{var}\left(\hat{\beta}_0\right) = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - $\text{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sigma^2 \frac{-\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} < 0$
  - Note that the covariance between the slope and intercept estimators is negative if $\bar{x} > 0$: overestimating one will tend to cause us to underestimate the other
- What determines the variance of $\hat{\beta}$?
  - Smaller variance of error $\Rightarrow$ more precise estimators
  - Larger number of observations $\Rightarrow$ more precise estimators
  - More dispersion of observations around mean $\Rightarrow$ more precise estimators
- What do we know about the overall **probability distribution of $\hat{\beta}$?**
  - If *u* is normal, then $\hat{\beta}$ is also normal because it is a linear function of the *u* variables and linear functions of normally distributed variables are also normally distributed.
  - If *u* is not normal, then $\hat{\beta}$ converges to a normal distribution as *n* $\rightarrow \infty$ provided some weak conditions on the distribution of *u* are satisfied.
    - In other words, the estimator is **asymptotically normal**

Estimating the variance of the estimator

- The expressions above are the **true variance/covariance** of the estimated coefficient vector. However, because we do not know σ^2, they are not of practical use to us. We need an estimator for σ^2 in order to calculate a **standard error** of the coefficients: an ***estimate*** **of their standard deviations**.
    - The required estimate in the classical case is $\hat{\sigma}^2 \equiv \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-2}$.
        - We divide by $n-2$ because this is the number of "degrees of freedom" in our regression.
        - Degrees of freedom are a very important issue in econometrics. It refers to how many data points are available *in excess of the minimum number required to estimate the model.*
        - In this case, it takes minimally two points to define a line, so the smallest possible number of observations for which we can fit a bivariate regression is 2. Any observations beyond 2 make it (generally) impossible to fit a line perfectly through all observations. Thus, $n-2$ is the number of degrees of freedom in the sample.
        - We always divide sums of squared residuals by the number of degrees of freedom in order to get unbiased variance estimates.
            - For example, in calculating the sample variance, we use $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})$ because there are $n-1$ degrees of freedom left after using one to calculate the mean.
            - Here, we have two coefficients to estimate, not just one, so we divide by $n-2$.
    - The *standard error* of each coefficient is the square root of the corresponding diagonal element of that estimated covariance matrix.

How good is the OLS estimator?

- Is OLS the best estimator? Under what conditions?
- Under "classical" regression assumptions SLR1–SLR5 (but not necessarily normality) the Gauss-Markov Theorem shows that the OLS estimator is BLUE.
    - Any other estimator that is unbiased and linear in u has higher variance than OLS.
    - Note that (5, 0) is an estimator with zero variance, but it is biased in the general case.
- Violation of any of the SLR1–SLR5 assumptions usually means that there is a better estimator.

Monte Carlo methods

Based on HGL Appendix 2G

- How do we evaluate an estimator such as OLS?
    - Under simple assumptions, we can sometimes calculate the estimator's theoretical probability distribution.
    - We can often calculate the theoretical distribution to which the estimator converges in large samples even when we cannot calculate the small-sample distribution.
    - In general (and, in particular, when we cannot calculate the true distribution), we can simulate the model over thousands of samples to estimate its distribution.
    - The estimation of the probability distribution of an estimator through simulation is called "Monte Carlo simulation" and is an increasingly important tool in econometrics.
    - Note similarity between Monte Carlo (theoretical distribution of u) and bootstrap (sample randomly among residuals)
- Consider simple Monte Carlo example: (MC Class Demo.dta)
    - Let's suppose that we are working with a given, fixed $n = 157$.
    - We have fixed, given values of the x variable for all 157 observations.
        - Using HGL's ex9-13.dta with advertising variable as x
    - We assume that the true population values of β_0 and β_1 are 10 and 3.
        - Close to estimated values for regression of sales on advertising
    - The true error term is IID normal with variance 0.09 (standard deviation 0.3)
- To use Monte Carlo to simulate the distribution of the OLS estimators, we generate M replications of the sampling experiment:
    - M sets of 157 IID $N(0, 0.09)$ simulated observations on u using random number generator
    - (We would generate sample values for x if it were not being taken as fixed.)
    - Calculate the M sets of 157 values of y_i for each observation as $\beta_1 + \beta_2 x_i + u_i$ with known values of the parameters and x and simulated values of u.
    - Run M regressions for the M simulated samples, keeping the estimated values of interest (presumably $\hat{\beta}_0$ and $\hat{\beta}_1$, but possibly also other values)
    - Look at distribution of the estimators over M replications to approximate the actual distribution
        - Mean
        - Variance/standard deviation/standard error
        - Quantiles for use in inference
    - Demonstrate using Stata

- Setup data
  - *x* is already in MC Class Demo.dta
- Create do file
  - program olstest
  - g u=rnormal(0, 0.3)
  - g y=10 + 3*x+u
  - reg y x
  - drop u y
  - end
- Load it into memory: run olstest
- Run simulation with 5000 replications
  - simulate b=_b[x] , reps(5000): olstest
- View data browser
- Summarize b
- Hist b
- Centile b , centile (2.5 97.5)
- Show summary stats, histogram, centiles (2.5, 97.5)

Least-squares regression model in matrix notation

(From Griffiths, Hill, and Judge, Section 5.4)

- We can write the *i*th observation of the bivariate linear regression model as
  $y_i = \beta_0 + \beta_1 x_i + u_i$.
- Arranging the *n* observations vertically gives us *n* such equations:
  $y_1 = \beta_0 + \beta_1 x_1 + u_1,$
  $y_2 = \beta_0 + \beta_1 x_2 + u_2,$
  $\vdots$
  $y_n = \beta_0 + \beta_1 x_n + u_n.$
- This is a system of linear equations that can be conveniently rewritten in matrix form. There is no real need for the matrix representation with only one regressor because the equations are simple, but when we add regressors the matrix notation is more useful.
  - Let **y** be an $n \times 1$ column vector:
    $$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

- Let $\mathbf{X}$ be an $n \times 2$ matrix:

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

- $\boldsymbol{\beta}$ is a 2×1 column vector of coefficients:

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

- And $\mathbf{u}$ is an $n \times 1$ vector of the error terms:

$$u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

- Then $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ expresses the system of N equations very compactly.
- (Write out matrices and show how multiplication works for single observation.)

- In matrix notation, $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\mathbf{b}$ is the vector of residuals.
- Summing squares of the elements of a column vector in matrix notation is just the inner product: $\sum_{i=1}^{n} \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}},$ where prime denotes matrix transpose. Thus we want to minimize this expression for least squares.

$$\begin{aligned} \hat{\mathbf{u}}'\hat{\mathbf{u}} &= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= (\mathbf{y}' - \mathbf{b}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}. \end{aligned}$$

- Differentiating with respect to the coefficient vector and setting to zero yields $-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0},$ or $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$.
- Pre-multiplying by the inverse of $\mathbf{X}'\mathbf{X}$ yields the OLS coefficient formula: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. (This is one of the few formulas that you need to memorize.)

- Note symmetry between matrix formula and scalar formula. $\mathbf{X}'\mathbf{y}$ is the sum of the cross product of the two variables and $\mathbf{X}'\mathbf{X}$ is the sum of squares of the regressor. The former is in the numerator (and not inverted) and the latter is in the denominator (and inverted).
- In matrix notation, we can express our estimator in terms of $\mathbf{u}$ as

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

- When we condition on *x*, the covariance matrix of the coefficient estimator is also easy to compute under the OLS assumptions.
  - **Covariance matrices**: The covariance of a vector random variable is a matrix with variances on the diagonal and covariances on the off-diagonals. For an $M \times 1$ vector random variable $\mathbf{z}$, the covariance matrix is to the following outer product:

$$\begin{aligned}\operatorname{cov}(\mathbf{z}) &= E\left(\left(\mathbf{z}-E\mathbf{z}\right)\left(\mathbf{z}-E\mathbf{z}\right)'\right) \\ &= \begin{pmatrix} E\left(z_1-Ez\right)^2 & E\left(z_1-Ez\right)\left(z_2-Ez\right) & \cdots & E\left(z_1-Ez\right)\left(z_M-Ez\right) \\ E\left(z_1-Ez\right)\left(z_2-Ez\right) & E\left(z_2-Ez\right)^2 & \cdots & E\left(z_2-Ez\right)\left(z_M-Ez\right) \\ \vdots & \vdots & \ddots & \vdots \\ E\left(z_1-Ez\right)\left(z_M-Ez\right) & E\left(z_2-Ez\right)\left(z_M-Ez\right) & \cdots & E\left(z_M-Ez\right)^2 \end{pmatrix}.\end{aligned}$$

  - In our regression model, if u is IID with mean zero and variance σ^2, then $E\mathbf{u} = 0$ and $\operatorname{cov}(\mathbf{u}) = E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}_n$, with $\mathbf{I}_n$ being the order-n identity matrix.
  - We can then compute the covariance matrix (conditional on x) of the (unbiased) estimator as

$$\begin{aligned}\operatorname{cov}\left(\hat{\beta}\right) &= E\left[\left(\hat{\beta}-\boldsymbol{\beta}\right)\left(\hat{\beta}-\boldsymbol{\beta}\right)'\right] \\ &= E\left[\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\right)\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\right)'\right] \\ &= E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E\left(\mathbf{u}\mathbf{u}'\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}.\end{aligned}$$

    - What happens to $\operatorname{var}\left(\hat{\beta}_1\right)$ as n gets large? Summations in $\mathbf{X}'\mathbf{X}$ have additional terms, so they get larger. This means that inverse matrix gets "smaller" and variance decreases: more observations implies more accurate estimators.
    - Note that variance also increases as the variance of the error term goes up. More imprecise fit implies less precise coefficient estimates.
  - Our *estimated* covariance matrix of the coefficients is then $\hat{\sigma}^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

- The (2, 2) element of this matrix is

$$\hat{\sigma}^2 \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{1}{n-2} \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- This is the formula we calculated in class for the scalar system.

- Thus, to summarize, when the classical assumptions hold and *e* is normally distributed, $\hat{\beta} \sim N\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$.

Asymptotic properties of OLS bivariate regression estimator

(Based on S&W, Chapter 17. Not covered in class Spring 2019)

- **Convergence in probability (probability limits)**
  - Assume that $S_1, S_2, \ldots, S_N, \ldots$ is a sequence of random variables.
    - In practice, they are going to be estimators based on 1, 2, …, *N* observations.
  - $S_N \xrightarrow{p} \mu$ if and only if $\lim_{N\to\infty} \Pr\left[|S_N - \mu| \geq \delta\right] = 0$ for any $\delta > 0$. Thus, for any small value of δ, we can make the probability that S_N is further from μ than δ arbitrarily small by choosing *N* large enough.
  - If $S_N \xrightarrow{p} \mu$, then we can write plim $S_N = \mu$.
  - This means that the entire probability distribution of S_N converges on the value μ as *N* gets large.
  - Estimators that converge in probability to the true parameter value are called **consistent estimators**.
- **Convergence in distribution**
  - If the sequence of random variables $\{S_N\}$ has cumulative probability distributions $F_1, F_2, \ldots, F_N, \ldots$, then $S_N \xrightarrow{d} S$ if and only if $\lim_{N\to\infty} F_N(t) = F(t)$, for all *t* at which *F* is continuous.
  - If a sequence of random variables converges in distribution to the normal distribution, it is called **asymptotically normal**.
- Properties of probability limits and convergence in distribution
  - Probability limits are very forgiving: Slutsky's Theorem states that
    - plim $(S_N + R_N)$ = plim S_N + plim R_N
    - plim $(S_N R_N)$ = plim $S_N \cdot$ plim R_N
    - plim (S_N / R_N) = plim S_N / plim R_N
  - The continuous-mapping theorem gives us
    - For continuous functions *g*, plim $g(S_N)$ = g(plim S_N)

        - And if $S_N \xrightarrow{d} S$, then $g(S_N) \xrightarrow{d} g(S)$.
    - Further, we can combine probability limits and convergence in distribution to get
        - If plim $a_N = a$ and $S_N \xrightarrow{d} S$, then
            - $a_N S_N \xrightarrow{d} aS$
            - $a_N \pm S_N \xrightarrow{d} a \pm S$
            - $S_N / a_N \xrightarrow{d} S / a$
        - These are *very* useful since it means that asymptotically we can treat any consistent estimator as a constant equal to the true value.
- **Central limit theorems**
    - There is a variety with slightly different conditions.
    - Basic result: If $\{S_N\}$ is a sequence of estimators of μ, then for a wide variety of underlying distributions, $\sqrt{N}(S_N - \mu) \xrightarrow{d} N(0, \sigma^2)$, where σ^2 is the variance of the underlying statistic.
- Applying asymptotic theory to the OLS model
    - Under the more general conditions than the ones that we have typically assumed (including, specifically, the finite kurtosis assumption, but not the homoskedasticity assumption or the assumption of fixed regressors), the OLS estimator satisfies the conditions for consistency and asymptotic normality.
    - $\sqrt{N}(b_2 - \beta_2) \xrightarrow{d} N\left(0, \frac{\text{var}\left[(x_i - E(x))e_i\right]}{\left[\text{var}(x_i)\right]^2}\right)$. This is general case with heteroskedasticity.
        - With homoskedasticity, the variable reduces to the usual formula:

$$\sqrt{N}(b_2 - \beta_2) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\left[\text{var}(x_i)\right]^2}\right).$$

    - $\text{plim}\,\hat{\sigma}^2_{b_2} = \sigma^2_{b_2}$, as proven in Section 17.3.
    - $t = \frac{b_2 - \beta_2}{s.e.(b_2)} \xrightarrow{d} N(0,1)$.
- Choice for *t* statistic:
    - If homoskedastic, normal error term, then exact distribution is t_{N-2}.
    - If heteroskedastic or non-normal error (with finite 4th moment), then exact distribution is unknown, but asymptotic distribution is normal
    - Which is more reasonable for any given application?