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.*
- Functional form
  - Is the relationship linear?
    - $y = \beta_1 + \beta_2 x$
    - This is natural first assumption, unless theory rejects it.
    - β_2 is slope, which determines whether relationship between x and y is positive or negative.
    - β_1 is 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
    - $E(y \mid x) = \beta_1 + \beta_2 x$: The *conditional mean* of y given x is $\beta_1 + \beta_2 x$
    - Adding an error term for a "stochastic" relationship gives us the actual value of y: $y = \beta_1 + \beta_2 x + e$
    - Error term e captures all of the above problems.
      - Error term is considered to be a random variable and is not observed directly.
      - Variance of e is σ^2, which is the *conditional variance* of y given x, the variance of the conditional distribution of y given x.
      - The simplest, but not usually valid, assumption is that the conditional variance is the same for all observations in our sample (*homoskedasticity*)

        - $\beta_2 = \frac{dE(y|x)}{dx}$, which means that the expected value of y increases by β_2 units when x increases by one unit
    - Does it matter which variable is on the left-hand side?
        - At one level, no:
            - $x = \frac{1}{\beta_2}(y - \beta_1 - e)$, so
            - $x = \gamma_1 + \gamma_2 y + v$, where $\gamma_1 \equiv -\frac{\beta_1}{\beta_2}, \quad \gamma_2 = \frac{1}{\beta_2}, \quad v = -\frac{1}{\beta_2}e.$
        - 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.
        - 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: $\text{cov}(y_i, y_j) = 0, \forall i \neq j.$
    - We will assume initially (for a few weeks) that the values of x are chosen as in an experiment: they are not random.
        - We will add random regressors soon and discover that they don't change things much as long as x is independent of e.
- Goals of regression
    - True regression line: actual relationship in population or DGP
        - True β and $f(e|x)$
        - Sample of observations comes from drawing random realizations of e from $f(e|x)$ 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_1 and β_2

- Estimate properties of probability distribution of error term *e*
- 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.
- SR1: $y = \beta_1 + \beta_2 x + e$
- SR2: $E(e) = 0$, so $E(y) = \beta_1 + \beta_2 x$
  - Note that if *x* is random, we make these conditional expectations
  - $E(e \mid x) = 0$
    $E(y \mid x) = \beta_1 + \beta_2 x$
- SR3: $\text{var}(e) = \sigma^2 = \text{var}(y)$
  - If *x* is random, this becomes $\text{var}(e \mid x) = \sigma^2 = \text{var}(y \mid x)$
  - We should (and will) consider the more general case in which variance varies across observations: *heteroskedasticity*
- SR4: $\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$
  - This, too, can be relaxed: *autocorrelation*
- SR5: *x* is non-random and takes on at least two values
  - We will allow random *x* later and see that $E(e \mid x) = 0$ implies that *e* must be uncorrelated with *x*.
- SR6: (optional) $e \sim N(0, \sigma^2)$
  - This is convenient, but not critical since the law of large numbers assures that for a wide variety of distributions of *e*, our estimators converge to normal as the sample gets large

Strategies for obtaining regression estimators

- What is an *estimator*?

- A rule (formula) for calculating an *estimate* of a parameter (β_1, β_2, or σ^2) based on the sample values y, x
- Estimators are often denoted by ^ over the variable being estimated: An estimator of β_2 might be denoted $\hat{\beta}_2$

- How might we estimate the β coefficients of the simple regression model?
  - Three strategies:
    - Method of least-squares
    - Method of moments
    - Method of maximum likelihood
  - All three strategies with the SR assumptions lead to the same estimator rule: the *ordinary least-squares* regression estimator: (b_1, b_2, s^2)
- **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 b_1, b_2, residuals are defined as $\hat{e}_i \equiv y_i - b_1 - b_2 x_i$
    - Or $\hat{e}_i = y_i - \hat{y}_i$, with $\hat{y}_i \equiv b_1 + b_2 x_i$
  - 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: $S = \sum_{i=1}^{N} \hat{e}_i^2 = \sum_{i=1}^{N} \left(y_i - b_1 - b_2 x_i \right)^2$
    - Least-squares estimators is the solution to $\min_{b_1, b_2} S$. Since S 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:**

- $$\frac{\partial S}{\partial b_2} = \sum_{i=1}^{N} 2\left(y_i - b_1 - b_2 x_i\right)\left(-x_i\right) = 0,$$
- $$-\sum_{i=1}^{N} y_i x_i + b_1 \sum_{i=1}^{N} x_i + b_2 \sum_{i=1}^{N} x_i^2 = 0.$$

$$\frac{\partial S}{\partial b_1} = \sum_{i=1}^{N} -2\left(y_i - b_1 - b_2 x_i\right) = 0$$

- $$\sum_{i=1}^{N} y_i - N b_1 - b_2 \sum_{i=1}^{N} x_i = 0$$

$$\overline{y} - b_1 - b_2 \overline{x} = 0$$

$$b_1 = \overline{y} - b_2 \overline{x}.$$

- Note that the b_1 condition assures that the regression line passes through the point $\left(\overline{x}, \overline{y}\right)$.
- Substituting the second condition into the first divided by N:

$$-\sum y_i x_i + \left(\overline{y} - b_2 \overline{x}\right) N\overline{x} + b_2 \sum x_i^2 = 0$$

$$-\left(\sum y_i x_i - N\overline{yx}\right) + b_2\left(\sum x_i^2 - N\overline{x}^2\right) = 0$$

$$b_2 = \frac{\sum y_i x_i - N\overline{yx}}{\sum x_i^2 - N\overline{x}^2} = \frac{\sum\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\sum\left(x_i - \overline{x}\right)^2} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}.$$

- The b_2 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 b_2.
  - 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 b_1.
  - 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.

- **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.

- Assume $E(e_i) = 0, \forall i$. 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 $\frac{1}{N}\sum \hat{e}_i = 0$. Thus we set the sample mean to the value we have assumed for the population mean.
- Assume $\text{cov}(x, e) = 0$, which is equivalent to $E\left[(x_i - E(x))e_i\right] = 0.$
  - Corresponding to this assumption about the population covariance between the regressor and the error term is the sample covariance condition: $\frac{1}{N}\sum (x_i - \bar{x})\hat{e}_i = 0.$ Again, we set the sample moment to the zero value that we have assumed for the population moment.

o Plugging the expression for the residual into the sample moment expressions above:

- $\frac{1}{N}\sum (y_i - b_1 - b_2 x_i) = 0,$
  $b_1 = \bar{y} - b_2 \bar{x}.$
- This is the same as the intercept estimate equation for the least-squares estimator above.
- $\frac{1}{N}\sum (x_i - \bar{x})(y_i - b_1 - b_2 x_i) = 0,$
  $\sum (x_i - \bar{x})(y_i - \bar{y} + b_2 \bar{x} - b_2 x_i) = 0,$
  $\sum (x_i - \bar{x})(y_i - \bar{y}) - \sum b_2 (x_i - \bar{x})(x_i - \bar{x}) = 0,$
  $b_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}.$
- This is exactly the same equation as for the OLS estimator.

o Thus, if we assume that $E(e_i) = 0, \forall i$ and $\text{cov}(x, e) = 0$ in the population, then the OLS estimator can be derived by the method of moments as well.

o (Note that both of these moment conditions follow from the extended assumption SR2 that $E(e|x) = 0$.)

- **Method of maximum likelihood**
  - o Consider the joint probability density function of y_i and x_i, $f_i(y_i, x_i | \beta_1, \beta_2)$. The function is written is conditional on the coefficients β to make explicit that the joint distribution of *y* and *x* are 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 probable 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 $\beta_1 = 0$ and $\beta_2 < 0$, then it is less likely that we would see observations where $y > 0$ when $x > 0$, than observations with $y < 0$.

- The idea of maximum-likelihood estimation is to choose a 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: $L_i(\beta_1,\beta_2 \mid x_i, y_i) \equiv f_i(x_i, y_i \mid \beta_1,\beta_2)$.
- 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:

$$f(x_1, y_1, x_2, y_2, \ldots, x_n, y_n \mid \beta_1,\beta_2) = \prod_{i=1}^{N} f_i(x_i, y_i \mid \beta_1,\beta_2)$$

- $$L(\beta_1,\beta_2 \mid x_1, y_1, x_2, y_2, \ldots, x_n, y_n) = \prod_{i=1}^{n} L_i(\beta_1,\beta_2 \mid x_i, y_i).$$

- If the conditional probability distribution of e conditional on x is Gaussian (normal) with mean zero and variance σ^2:
  - $f_i(x_i, y_i \mid \beta_1,\beta_2) = L_i(\beta_1,\beta_2 \mid x_i, y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(\frac{-\frac{1}{2}(y_i-\beta_1-\beta_2 x_i)^2}{\sigma^2}\right)}$
  - 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: $\ln L_i = -\frac{1}{2}\ln(\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - \beta_1 - \beta_2 x_i)^2$
  - Aggregating over the sample:

$$\ln \prod_{i=1}^{N} L_i(\beta_1,\beta_2 \mid x_i, y_i) = \sum_{i=1}^{N} \ln L_i(\beta_1,\beta_2 \mid x_i, y_i)$$
$$= \sum_{i=1}^{N} \left[-\frac{1}{2}\ln(\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - \beta_1 - \beta_2 x_i)^2 \right]$$
$$= -\frac{N}{2}\ln(\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_i)^2.$$

    - The only part of this expression that depends on β 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 *e* conditional on *x* is Gaussian with mean zero and constant variance σ^2, and if the observations are IID.
- 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.

Sampling distribution of OLS estimators

- b_1 and b_2 are random variables: they are functions of the random variables *y* and *e*.
    - We can think of the probability distribution of *b* 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. (See below)
    - Because OLS estimator (under our assumptions) is linear, we can derive its distribution.

- We can write the OLS slope estimator as

$$b_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{N}(x_i - \bar{x})^2}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N}(\beta_1 + \beta_2 x_i + e_i - \bar{y})(x_i - \bar{x})}{\frac{1}{N}(x_i - \bar{x})^2}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N}(\beta_1 + \beta_2 x_i + e_i - (\beta_1 + \beta_2 \bar{x}))(x_i - \bar{x})}{\frac{1}{N}(x_i - \bar{x})^2}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N}(\beta_2 (x_i - \bar{x}) + e_i)(x_i - \bar{x})}{\frac{1}{N}(x_i - \bar{x})^2}$$

$$= \beta_2 + \frac{\frac{1}{N}\sum_{i=1}^{N} e_i (x_i - \bar{x})}{\frac{1}{N}(x_i - \bar{x})^2}$$

  The third step uses the property $\bar{y} = \beta_1 + \beta_2 \bar{x}$, since the expected value of *e* is zero.

- For now, we are assuming that ***x* is non-random**, as in a controlled experiment.
  - If *x* is fixed, then the only part of the formula above that is random is *e*.
  - The formula shows that the slope estimate is linear in *e*.
  - This means that if *e* is Gaussian, then the slope estimate will also be Gaussian.
    - Even if *e* is not Gaussian, the slope estimate will converge to a Gaussian distribution as long as some modest assumptions about its distribution are satisfied.
  - Because all the *x* variables are non-random, they can come outside when we take expectations, so

$$E(b_2) = \beta_2 + E\left[\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})e_i}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}\right] = \beta_2 + \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})E(e_i)}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \beta_2 .$$

  - What about the **variance** of b_2?
    - We will do the details of the analytical work in matrix form because it's easier

- $$\begin{aligned}\operatorname{var}(b_2) &= E(b_2-\beta_2)^2 \\ &= E\left[\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})E(e_i)}{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}\right]^2 \\ &= \dots \\ &= \frac{\sigma^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}.\end{aligned}$$
  - HGL equations 2.14 and 2.16 provide formulas for variance of b_1 and the covariance between the coefficients:
    - $\operatorname{var}(b_1) = \sigma^2 \dfrac{\sum_{i=1}^{N} x_i^2}{N\sum_{i=1}^{N}(x_i-\bar{x})^2}$
    - $\operatorname{cov}(b_1, b_2) = \sigma^2 \dfrac{-\bar{x}}{\sum_{i=1}^{N}(x_i-\bar{x})^2} < 0$
      - Note that the covariance between the slope and intercept estimators is negative: overestimating one will tend to cause us to underestimate the other
  - What determines the variance of *b*?
    - Smaller variance of error ⇒ more precise estimators
    - Larger number of observations ⇒ more precise estimators
    - More dispersion of observations around mean ⇒ more precise estimators
  - What do we know about the overall **probability distribution of *b*?**
    - If assumption SR6 is satisfied and *e* is normal, then *b* is also normal because it is a linear function of the *e* variables and linear functions of normally distributed variables are also normally distributed.
    - If assumption SR6 is not satisfied, then *b* converges to a normal distribution as $N \to \infty$ provided some weak conditions on the distribution of *e* are satisfied.
- These expressions are the **true variance/covariance** of the estimated coefficient vector. However, because we do not know σ^2, it is 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 deviation.

- The required estimate in the classical case is $s^2 \equiv \frac{1}{N-2}\sum_{i=1}^{N}\hat{e}_i^2$.
    - 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.
- Note that the HGL text uses an alternative formula based on $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\hat{e}_i^2$.
    - This estimator for σ^2 is biased because there are only $N-2$ degrees of freedom in the N residuals—2 are used up in estimating the 2 β parameters.
    - In large samples they are equivalent.

How good is the OLS estimator?

- Is OLS the best estimator? Under what conditions?
- Under "classical" regression assumptions SR1–SR5 (but not necessarily SR6) the Gauss-Markov Theorem shows that the OLS estimator is BLUE.
    - Any other estimator that is unbiased and linear in *e* has higher variance than *b*.
    - Note that (5, 0) is an estimator with zero variance, but it is biased in the general case.
- Violation of any of the SR1–SR5 assumptions usually means that there is a better estimator.

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
  $y_i = \beta_1 + \beta_2 x_i + e_i$.
- Arranging the n observations vertically gives us n such equations:
  $y_1 = \beta_1 + \beta_2 x_1 + e_1$,
  $y_2 = \beta_1 + \beta_2 x_2 + e_2$,
  $\vdots$
  $y_N = \beta_1 + \beta_2 x_N + e_N$.
- 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 $\mathbf{y}$ 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:
    $$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$
  - And $\mathbf{e}$ is an $n \times 1$ vector of the error terms:
    $$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{pmatrix}.$$
  - Then $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ expresses the system of N equations very compactly.
  - (Write out matrices and show how multiplication works for single observation.)
- In matrix notation, $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{Xb}$ is the vector of residuals.

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

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

  - 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: $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. (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 $\mathbf{e}$ as

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

  - When x is non-stochastic, 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:

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

    - In our regression model, if e is IID with mean zero and variance σ^2, then $E\mathbf{e} = 0$ and $\operatorname{cov}(\mathbf{e}) = E(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I}_N$, with $\mathbf{I}_N$ being the order-N identity matrix.

- We can then compute the covariance matrix of the (unbiased) estimator as

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

  - What happens to $\operatorname{var}(b_i)$ 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 $s^2(\mathbf{X}'\mathbf{X})^{-1}$.
  - The (2, 2) element of this matrix is

$$s^2\frac{1}{\sum_{i=1}^{N}(x_i-\bar{x})^2} = \frac{1}{N-2}\frac{\sum_{i=1}^{N}\hat{e}_i^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}.$$

  - 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, $\mathbf{b} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$.

Asymptotic properties of OLS bivariate regression estimator

(Based on S&W, Chapter 17.)

- **Convergence in probability (probability limits)**
  - Assume that S_1, S_2, …, S_N, … 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 $\lim_{N\to\infty} \Pr\left[\left|S_N - \mu\right| \geq \delta\right] = 0$ 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 $g(S_N) = g(\text{plim } S_N)$
    - 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.

- $\sqrt{N}\left(b_2-\beta_2\right)\xrightarrow{d}N\left(0,\frac{\text{var}\left[\left(x_i-E(x)\right)e_i\right]}{\left[\text{var}\left(x_i\right)\right]^2}\right)$. This is general case with heteroskedasticity.
  - With homoskedasticity, the variable reduces to the usual formula: $\sqrt{N}\left(b_2-\beta_2\right)\xrightarrow{d}N\left(0,\frac{\sigma^2}{\left[\text{var}\left(x_i\right)\right]^2}\right)$.
- $\text{plim}\,\hat{\sigma}^2_{b_2}=\sigma^2_{b_2}$, as proven in Section 17.3.
- $t=\frac{b_2-\beta_2}{s.e.\left(b_2\right)}\xrightarrow{d}N\left(0,1\right)$.

- 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?

Linearity and nonlinearity

- The OLS estimator is a linear estimator because *b* is linear in *e* (which is because *y* is linear in β), not because *y* is linear in *x*.
- OLS can easily handle nonlinear relationships between *y* and *x*.
  - $\ln y = \beta_1 + \beta_2 x$
  - $y = \beta_1 + \beta_2 x^2$
  - etc.
- Dummy (indicator) variables take the value zero or one.
  - Example: MALE = 1 if male and 0 if female.
  - $y_i=\beta_1+\beta_2 MALE_i+e_i$
    - For females, $E\left[y\,|\,MALE\right]=\beta_1$
    - For males, $E\left[y\,|\,MALE\right]=\beta_1+\beta_2$
    - Thus, β_2 is the difference between the expected value of males and females.