# Section 2 Simple Regression

## What regression does

- Relationship
  - 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.
- Form
  - Is the relationship linear?
    - $Y = \beta_0 + \beta_1 X$
    - This is natural first assumption, unless theory rejects it.
    - $\beta_1$ is slope, which determines whether relationship between $X$ and $Y$ is positive or negative.
    - $\beta_0$ 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
  - Adding an error term for a "stochastic" relationship
    - $Y = \beta_0 + \beta_1 X + u$
    - Error term $u$ captures all of the above problems.
    - Error term is considered to be a random variable and is not observed directly.
  - 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 $\gamma_0 \equiv -\frac{\beta_0}{\beta_1}, \quad \gamma_1 = \frac{1}{\beta_1}, \quad v = -\frac{1}{\beta_1} u.$
    - 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.
    - 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
- Goals of regression
    - True regression line: actual relationship in population or DGP
        - True $\beta$ 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 $\beta_0$ and $\beta_1$
        - 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

# Strategies for obtaining regression estimators

- How might we estimate the $\beta$ coefficients of the simple regression model? Three strategies that all lead (in this case and under appropriate assumptions) to the same result:
    - Method of least-squares
    - Method of moments
    - Method of maximum likelihood
- **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 $\hat{u}_i \equiv Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$
    - Why not use 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: $S = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left( Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)^2$
  - Least-squares estimators is the solution to $\min_{\hat{\beta}_0, \hat{\beta}_1} S$. 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:**
    - $$\frac{\partial S}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2\left( Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right)\left( -X_i \right) = 0,$$
    $$-\sum_{i=1}^{n} Y_i X_i + \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0.$$
    $$\frac{\partial S}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} -2\left( Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right) = 0$$
    - $$\sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0$$
    $$\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0$$
    $$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$
  - Note that the $\beta_0$ condition assures that the regression line passes through the point $\left( \bar{X}, \bar{Y} \right)$.

- Substituting the second condition into the first divided by $n$:

$$-\sum Y_i X_i + \left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) n\bar{X} + \hat{\beta}_1 \sum X_i^2 = 0$$

$$-\left(\sum Y_i X_i - n\bar{Y}\bar{X}\right) + \hat{\beta}_1 \left(\sum X_i^2 - n\bar{X}^2\right) = 0$$

$$\hat{\beta}_1 = \frac{\sum Y_i X_i - n\bar{Y}\bar{X}}{\sum X_i^2 - n\bar{X}^2} = \frac{\sum \left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\sum \left(X_i - \bar{X}\right)^2} = \frac{s_{XY}}{s_X^2}.$$

- The $\beta_1$ estimate is the sample covariance of $X$ and $Y$ divided by the sample variables 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 $\hat{\beta}_0$.
  - 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.

o 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 $E\left(u_i\right) = 0, \forall i$. This means that the population/DGP mean of the error term is zero.
      - Corresponding to this assumption about the population mean of $u$ 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.
  - $\frac{1}{n}\sum\left(X_i - \bar{X}\right)\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0,$
    $\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i\right) = 0,$
  - $\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) - \sum \hat{\beta}_1 \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right) = 0,$
    $\hat{\beta}_1 = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum\left(X_i - \bar{X}\right)^2}.$
    - This is exactly the same equation as for the OLS estimator.
- Thus, if we assume that $E\left(u_i\right) = 0, \forall i$ and $\text{cov}\left(X, u\right) = 0$ in the population, then the OLS estimator can be derived by the method of moments as well.
- (Note that both of these moment conditions follow from S&W's assumption that $E(u \mid X) = 0$.)

- **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 is conditional on the coefficients $\beta$ 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 $\beta_0 = 0$ and $\beta_1 < 0$, 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: $L_i\left(\beta_0, \beta_1 \mid X_i, Y_i\right) \equiv f_i\left(X_i, Y_i \mid \beta_0, \beta_1\right).$
  - 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:

- $$f\left(X_1,Y_1,X_2,Y_2,\ldots,X_n,Y_n \mid \beta_0,\beta_1\right)=\prod_{i=1}^{n} f_i\left(X_i,Y_i \mid \beta_0,\beta_1\right)$$
- $$L\left(\beta_0,\beta_1 \mid X_1,Y_1,X_2,Y_2,\ldots,X_n,Y_n\right)=\prod_{i=1}^{n} L_i\left(\beta_0,\beta_1 \mid X_i,Y_i\right).$$

  - If the conditional probability distribution of *u* conditional on *X* is Gaussian (normal) with mean zero and variance $\sigma^2$:
    - $f_i\left(X_i,Y_i \mid \beta_0,\beta_1\right)=L_i\left(\beta_0,\beta_1 \mid X_i,Y_i\right)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{\left(\frac{-\frac{1}{2}\left(Y_i-\beta_0-\beta_1X_i\right)^2}{\sigma^2}\right)}$
    - 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: $\ln L_i=-\frac{1}{2}\ln\left(\pi\sigma^2\right)-\frac{1}{2\sigma^2}\left(Y_i-\beta_0-\beta_1X_i\right)^2$
    - Aggregating over the sample:

$$\ln\prod_{i=1}^{n} L_i\left(\beta_0,\beta_1 \mid X_i,Y_i\right)=\sum_{i=1}^{n}\ln L_i\left(\beta_0,\beta_1 \mid X_i,Y_i\right)$$
$$=\sum_{i=1}^{n}\left[-\frac{1}{2}\ln\left(\pi\sigma^2\right)-\frac{1}{2\sigma^2}\left(Y_i-\beta_0-\beta_1X_i\right)^2\right]$$
$$=-\frac{n}{2}\ln\left(\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_i\right)^2.$$

  - 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 *u* conditional on *X* is Gaussian with mean zero and constant variance $\sigma^2$, 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.

# 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:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + u_1, \\ Y_2 &= \beta_0 + \beta_1 X_2 + u_2, \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n + u_n. \end{aligned}$$

- 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:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

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

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}.$$

    - $\beta$ is a $2 \times 1$ column vector of coefficients:

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

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

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

    - Then $Y = X\beta + u$ expresses the system of $n$ equations very compactly.

- (Write out matrices and show how multiplication works for single observation.)

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

$$\begin{aligned}\hat{u}'\hat{u} &= \left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right) \\ &= \left(Y' - \hat{\beta}'X'\right)\left(Y - X\hat{\beta}\right) \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}.\end{aligned}$$

  - Differentiating with respect to the coefficient vector and setting to zero yields $-2X'Y + 2X'X\hat{\beta} = 0,$ or $X'X\hat{\beta} = X'Y.$
  - Pre-multiplying by the inverse of $X'X$ yields the OLS coefficient formula: $\hat{\beta} = \left(X'X\right)^{-1} X'Y.$ (This is one of the few formulas that you need to memorize.)
- Note symmetry between matrix formula and scalar formula. $X'Y$ is the sum of the cross product of the two variables and $X'X$ 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).

# Assumptions of OLS regression

The assumptions necessary to justify the use of OLS depend a lot on what you are using it for. Different texts are aiming at different levels of appropriateness, so they vary somewhat in the assumptions that they list.

- OLS is an estimator selected by the method of least squares and method of moments regardless of the underlying model (as long as the relevant moments exist).
- With Gaussian IID errors that are independent of $X$, OLS is MLE.
- **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.
- Stock and Watson's assumptions:
  - **Assumption #1**: $E\left(u_i \mid X_i\right) = 0.$
    - This implies two conditions that are often expressed separately:

- $E(u_i) = 0$
- $\text{cov}(X_i, u_i) = 0.$

- In many ways, this is the most crucial of the assumptions.
  - Regression takes whatever correlation exists between $Y$ and $X$ to be the regression relationship. If $u$ moves together (in either direction) with $X$, then part of the observed correlation between $Y$ and $X$ will be the effect of $u$, which will be incorrectly captured as the effect of $X$.
  - This will lead to bias in the estimate of $\beta_1$, which we will demonstrate later. Because $u$ includes the effects of variables that are not included in the regression, this often results from "omitted-variable bias."
- This assumption is fulfilled most obviously if $X$ is truly exogenous as in a random controlled experiment.
  - Because this is uncommon in economics, much of econometrics (as distinct from statistics as applied to biology or psychology) revolves around how to deal with data not generated from controlled experiments.
- As shown in the method of moments, the OLS estimators are constructed in a way that assures that
  - $\sum_{i=1}^{n} \hat{u}_i = 0$ ,
  - $\sum_{i=1}^{n} X_i \hat{u}_i = 0$ .
  - Because these conditions are imposed in the calculation of OLS estimators, we can't use them as the basis for testing whether $E(X_i \mid u_i) = 0$ .

- **Assumption #2**: The observations ($X_i$, $Y_i$) in the sample are IID.
  - This is often characterized as $u_i$ being IID with $X$ fixed.
  - Key implications of this assumption:
    - Error terms of all observations are uncorrelated: *absence of autocorrelation*.
    - Special case of Assumption #0: Same model applies to all observations.
  - Time-series data are almost always autocorrelated.
  - Cross-section data may be spatially autocorrelated.
- **Assumption #3**: Large outliers are unlikely.
  - Finite kurtosis.

- This assumption allows us to use central-limit theorems to argue that the probability distributions of test statistics such as the OLS coefficient estimators converge to normal distributions with large samples.
- In practice, this assumption helps assure that huge outliers do not drag regression line away from the majority of points. (See S&W's Figure 4.5 on page 130.)

# Sampling distribution of OLS estimators

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables: they are functions of the random variables *Y*, *X*, and *u*.
  - We can think of the probability distribution of $\hat{\beta}$ as occurring over repeated random samples from the underlying population or DGP.
- Following methods shown in S&W Appendix 4.3, we can write the OLS slope estimator as

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

- In matrix notation, the corresponding algebra is

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

- A simple, illustrative special case is when ***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 *u*.
  - The formula shows that the slope estimate is linear in *u*.
  - This means that if *u* is Gaussian, then the slope estimate will also be Gaussian.
  - Because all the *X* variables are non-random, they can come outside when we take expectations, so

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

  - 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 $z$, the covariance matrix is to the following outer product:

$$\operatorname{cov}(z) = E\left((z - Ez)(z - Ez)'\right)$$

$$= \begin{pmatrix} E(z_1 - Ez)^2 & E(z_1 - Ez)(z_2 - Ez) & \dots & E(z_1 - Ez)(z_m - Ez) \\ E(z_1 - Ez)(z_2 - Ez) & E(z_2 - Ez)^2 & \dots & E(z_2 - Ez)(z_m - Ez) \\ \vdots & \vdots & \ddots & \vdots \\ E(z_1 - Ez)(z_m - Ez) & E(z_2 - Ez)(z_m - Ez) & \dots & E(z_m - Ez)^2 \end{pmatrix}.$$

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

$$\operatorname{cov}(\hat{\beta}) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right]$$

$$= E\left[\left((X'X)^{-1}X'u\right)\left((X'X)^{-1}X'u\right)'\right]$$

$$= E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}\right]$$

$$= (X'X)^{-1}X'E(uu')X(X'X)^{-1}$$

$$= \sigma_u^2 (X'X)^{-1}X'X(X'X)^{-1} = \sigma_u^2 (X'X)^{-1}.$$

  - What happens to $\operatorname{var}(\hat{\beta}_i)$ as $n$ gets large? Summations in $X'X$ 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.
- This expression is the *true variance/covariance* of the estimated coefficient vector. However, because we do not know $\sigma_u^2$, it is not of practical use to us. We need an estimator for $\sigma_u^2$ in order to calculate a **standard error** of the coefficients: an estimate of their standard deviation.
  - The required estimate in the classical case is $s_{\hat{u}}^2 \equiv \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2$.

- Our estimated covariance matrix of the coefficients is then $s_{\hat{u}}^2 \left( X'X \right)^{-1}$. The (2, 2) element of this matrix is

$$s_{\hat{u}}^2 \frac{1}{\sum_{i=1}^{n} \left( X_i - \bar{X} \right)^2} = \frac{1}{n-2} \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n} \left( X_i - \bar{X} \right)^2}.$$

- The standard error of each coefficient is the square root of the corresponding diagonal element of that estimated covariance matrix.

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

- In the general case, where the OLS assumptions hold but ***X* is a random variable**:
  - OLS estimator is still unbiased as long as Assumptions #1 and #2 hold.
  - Equation (4.31) on page 145 of Appendix 4.3 does this math using the Law of Iterated Expectations: $E(Z) = E\left[ E\left( Z \mid X \right) \right]$, but with some confusing missed steps.
  - Starting out with the Law of Iterated Expectations,

$$E\left( \hat{\beta}_1 \right) = E\left[ E\left( \hat{\beta}_1 \mid X_1, \ldots, X_n \right) \right]$$

$$= \beta_1 + E\left[ E\left( \left[ \frac{\frac{1}{n} \sum_{i=1}^{n} \left( X_i - \bar{X} \right) u_i}{\frac{1}{n} \sum_{i=1}^{n} \left( X_i - \bar{X} \right)^2} \right] \mid X_1, \ldots, X_n \right) \right].$$

    Now, because the expectation in question is *conditional on X*, we can treat *X* as fixed and pull it outside the expectations operator:

$$E\left( \hat{\beta}_1 \right) = \beta_1 + E\left[ \frac{\frac{1}{n} \sum_{i=1}^{n} \left( X_i - \bar{X} \right) E\left( u_i \mid X_1, \ldots, X_n \right)}{\frac{1}{n} \sum_{i=1}^{n} \left( X_i - \bar{X} \right)^2} \right].$$

    From this point, S&W's argument follows directly: $E\left( u_i \mid X_1, \ldots, X_n \right) = 0$ because $u_i$ is independent both of all other observations (by Assumption #2) and of $X_i$ through Assumption #1's condition that $E\left( u_i \mid X_i \right) = 0$.

  - Thus, the OLS slope estimator is unbiased with stochastic *X* (as long as Assumptions #1 and #2 hold).
  - The variance of $\hat{\beta}$ in the stochastic-regressors case is more complicated, but in large samples we can determine the results.
    - S&W present scalar equations on page 133:

- $\text{var}\left(\hat{\beta}_1\right) = \frac{1}{n}\frac{\text{var}\left[\left(X_i - EX\right)u_i\right]}{\left[\text{var}\left(X_i\right)\right]^2}$.
    - What happens to this variance as *n* gets large? Fraction in front gets small, so again the estimator is more precise with more (independent) observations.
    - As before, an increase in the variance of *u* also raises variance of estimates through larger numerator.
- In order to use this variance in inference, we will need to estimate it to get the standard error of the coefficient estimate.
    - S&W present the following formula for the estimated variance of the slope estimator:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n}\frac{\frac{1}{n-2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]^2} = \frac{n}{n-2}\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \hat{u}_i^2}{\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right]^2}.$$

    - Compare this formula to the classical OLS standard error formula above: $\frac{1}{n-2}\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$. In the more general formula, squared residuals are weighted by the squared deviation of the regressor from its mean, giving more weight to the residuals associated with observations farther from the mean of *X*. In the classical formula, they are equally weighted.

# How good is the OLS estimator?

(Based on S&W 5.4 and 5.5.)

- Is OLS the best estimator? Under what conditions?
- "Classical" regression assumptions
    - S&W's assumptions 1-3, plus *homoskedasticity*: $\text{var}\left(u_i \mid X_1, X_2, \ldots, X_n\right) = \sigma_u^2$.
        - (Isn't homoskedasticity implied if *Y* and *X* are jointly IID? How can the variance of the error term differ across observations if the variance of *X* and the variance of *Y* (and their covariance) are identical and if all the observations follow the same model?)
    - Under these assumptions, the Gauss-Markov Theorem shows that the OLS estimator is BLUE.
        - Derivation above that showed that $\text{cov}\hat{\beta} = \sigma_u^2\left(X'X\right)^{-1}$ relied on the assumption of homoskedasticity.

        - Stata assumes homoskedasticity and calculates standard errors according to this formula, which is only correct when the error is homoskedastic.
- Homoskedasticity is often violated, so we can't often rely on OLS being BLUE.
    - Implications of heteroskedasticity (which is an example of a violation of Assumption #0):
        - OLS coefficient estimators are still unbiased and asymptotically normal.
        - OLS estimators are not efficient: there is a better estimator (WLS) in the presence of heteroskedasticity.
        - The "standard" Stata standard errors are incorrect.
        - Use "robust" option to get standard errors calculated with an estimator that does not assume homoskedasticity. (We'll study this in section 18.2.)
    - Testing for heteroskedasticity
        - Is the variance of the error term constant, random, or related to something?
        - Residuals are our best estimate of error term and, with mean zero, $E\left(\hat{u}_i^2\right) = \text{var}\left(u_i\right)$.
        - So we might want to look at whether the squared residuals are constant or if they vary systematically.
        - Of course, the actual squared residuals will vary randomly around their expected value(s) even if the actual underlying expected values (variance of the error terms) does not vary.
            - Thus, there is no general test for random heteroskedasticity. (We can't test the general hypothesis that var($u_i$) is a constant against the alternative that it varies according to some unspecified pattern.)
            - However, if var($u_i$) is related to some variable $Z_i$, then we would expect that $\hat{u}_i^2$ would also be related to $Z_i$.
            - If we have some idea what variable $Z$ might be related to the variance, we can regress the squared residuals on $Z$ and use the $t$ test to determine whether the relationship is significant using the methods we will soon discuss for testing hypotheses about regression coefficients.
            - Breusch-Pagan test: Regress squared residuals on regressors.
            - White test: Regress squared residuals on all regressors, squared regressors, and cross-products of regressors and test joint null hypothesis that all coefficients are zero. (This is multiple regression, which we don't know how to do yet and can get very large if there are lots of regressors.) Alternatively, you can regress squared residuals on fitted values and their squares.

- Goldfeld-Quandt test: If we believe that the variance might be different across different sub-samples, we can test whether the mean of $\hat{u}_i^2$ differs across the sub-samples. (This is the same as regressing $\hat{u}_i^2$ on a set of dummy variables corresponding to all but one of the sub-samples.)

# Measuring goodness of fit

- It is always of interest to have a measure of how well our regression line fits the data. There are several that are commonly reported.
- SSR = $\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2$, with $\hat{Y}_i \equiv Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$.
- TSS = total sum of squares = $\sum_{i=1}^{n} \left( Y_i - \bar{Y} \right)^2$.
- ESS = explained sum of squares = $\sum_{i=1}^{n} \left( \hat{Y}_i - \bar{Y} \right)^2$
    - Warning about notation: some books use RSS for regression sum of squares and ESS to mean error sum of squares.
- Fundamental regression identity: TSS = ESS + SSR. Works due to the enforced independence of $\hat{Y}$ and $\hat{u}$.
- **Standard error of the regression (estimate)**: This is our estimate of the standard deviation of the error term.
    - $s_{\hat{u}} = \sqrt{s_{\hat{u}}^2} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2} = \sqrt{\frac{SSR}{n-2}}$.
    - 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 $n-2$.

  - Standard error of regression is often (as in Stata) called *root mean squared error* or RMSE.

- **Coefficient of determination: $R^2$**
  - The $R^2$ coefficient measures the fraction of the variance in the dependent variable that is explained by the covariation with the regressor. It has a range of (0, 1), with $R^2 = 0$ meaning no relationship and $R^2 = 1$ meaning a perfect linear fit.
  - $R^2 \equiv \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = 1 - \frac{n-2}{n-1}\frac{s_{\hat{u}}^2}{s_Y^2}$.