PS2030
Political Research and Analysis
Unit 1: Fundamentals of Linear Regression
2. Assumptions of OLS Regression
3. Hypothesis Testing

Spring 2025

WW Posvar Hall 3600

Professor Steven Finkel

Plan for Sessions

- Assumptions of OLS regression: what they are, why they are needed, what they provide for us
- Hypothesis testing in bivariate regression models
  - Is the slope "statistically significant"?
  - Does the equation as a whole explain a "significant" amount of variation in Y?
- Confidence intervals for regression coefficients

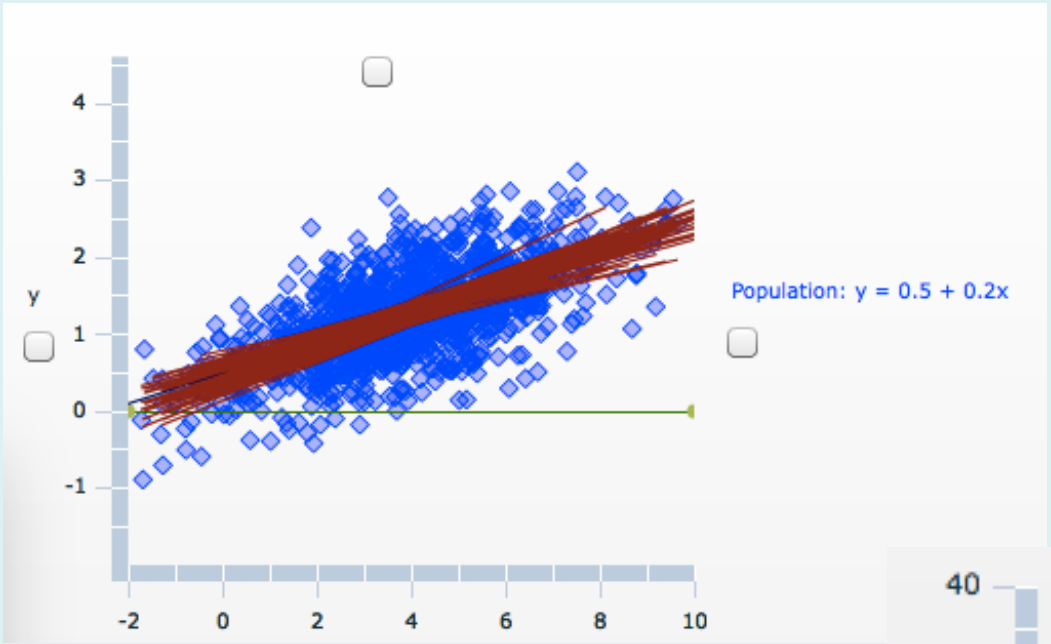
2. The Assumptions of OLS Regression

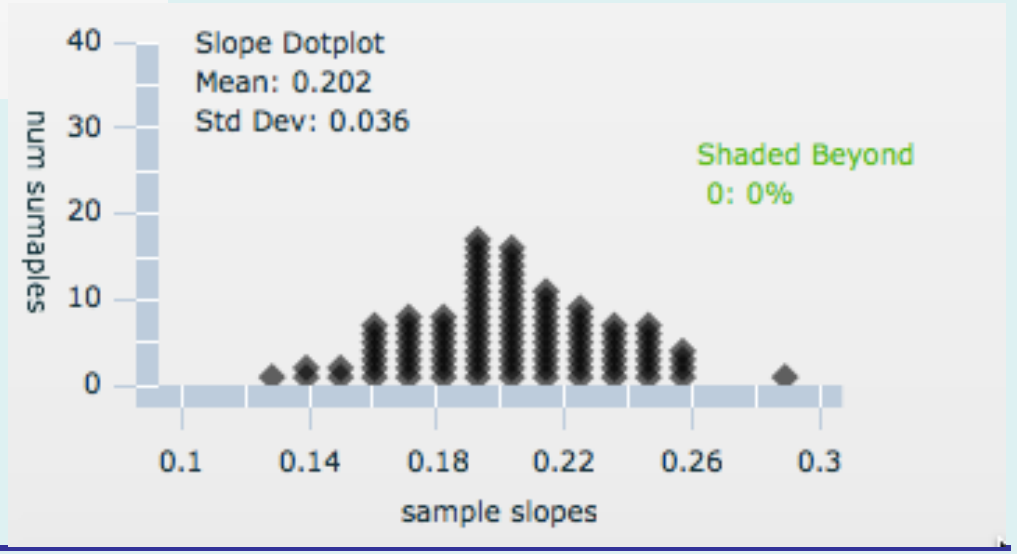
- We want to use the estimates from the SRF to make inferences about the PRF
- But we can only "trust" the estimates produced by OLS when certain assumptions about the population and the PRF are true
- OLS is one of many possible "estimators" that could be used to arrive at "estimates" of population parameters. Previous slides showed that it is based on intuitive logic and that the OLS regression line has many desirable properties. **But it will give us inaccurate information about the PRF unless certain assumptions hold.**
- What do we mean by "inaccurate information"?
- How do we know if the assumptions hold, given that population distributions, relationships, and parameters governing the relationships are almost always unknown?
- What should we do about it if the assumptions don't hold?

- The goal of providing "accurate information" about the PRF can be recast in terms of what we want to see from the ***sampling distribution*** of SRF estimates that is produced by a given estimator
- The sampling distribution of *slopes* is what would result if we were to sample the population an infinite number of times, calculate the slope using a given estimator, and plot the distribution of the sample slopes
- We use this sampling distribution to make inferences about the population slope from the slope that we observe in our **single** sample

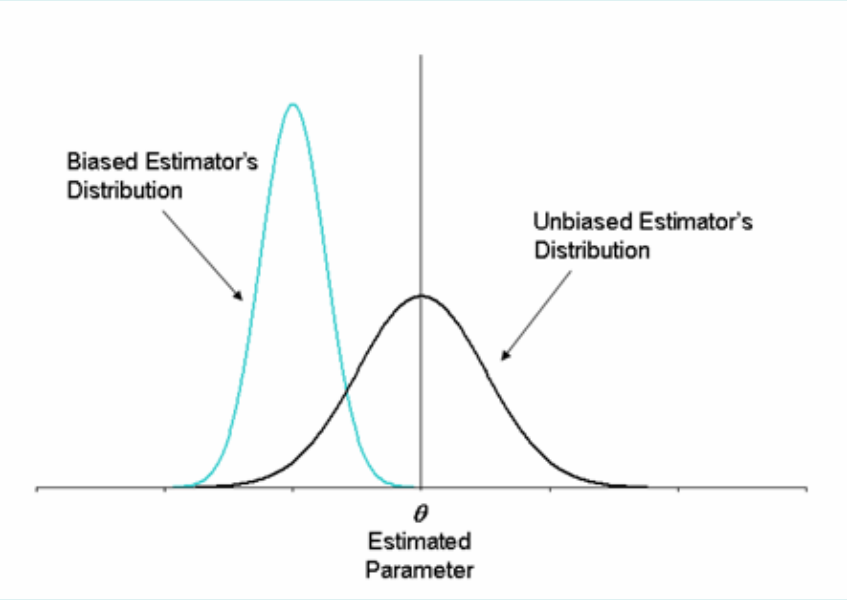
- **We can say that the assumptions of OLS need to be satisfied in order to ensure that the sampling distribution of slopes that is produced by the OLS procedure can be used to make accurate inferences from our single sample to the population parameters**

Example of a simulated OLS "empirical" sampling distribution: N=50, β_0=.5 β_1=.2, σ=.5, number of samples=100



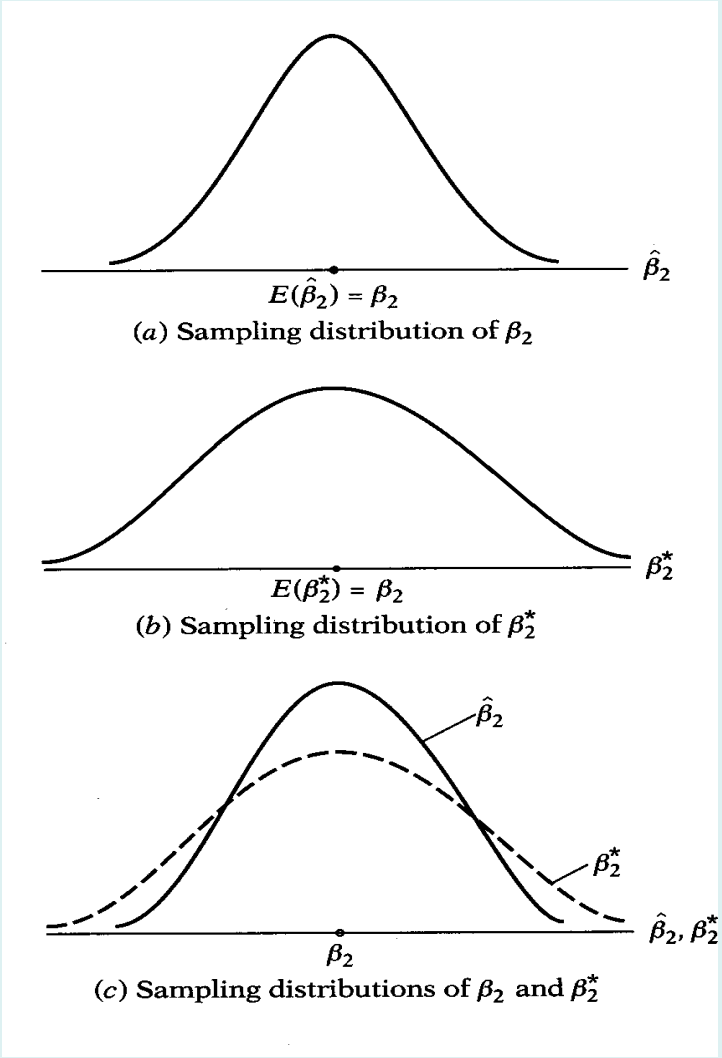


Desirable Properties of Estimators: Unbiasedness



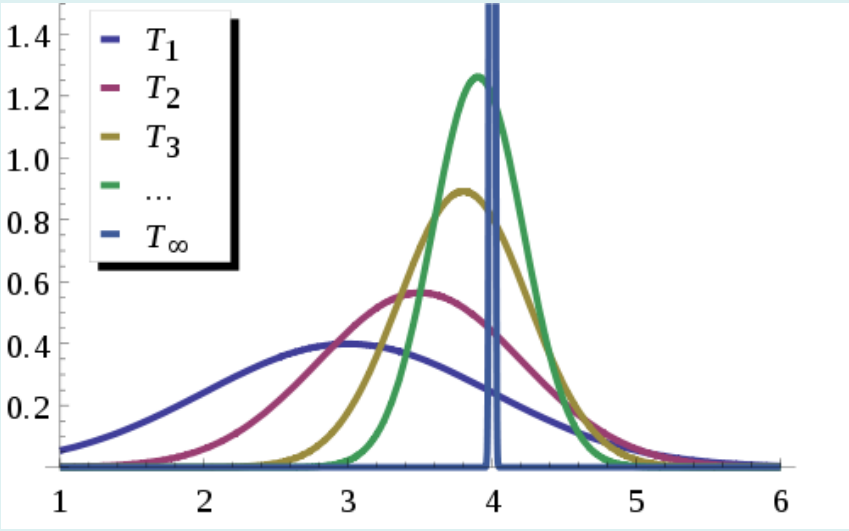
- We want the sampling distribution of an estimator to be centered around the true population parameter
- E ($\hat{\beta}$)=β
- That is, on average we want the value of the slope we estimate with our sample data to be the true population value; we neither underestimate nor overestimate the population value using the given estimator.
- OLS (and many other estimators in regression) will be biased under some conditions!

Desirable Properties of Estimators: Efficiency



- Estimators produce sampling distributions with different variances – some produce estimates that are very similar from sample to sample, and others produce estimates that vary widely (given fixed population characteristics)
- Other things being equal, estimators with smaller variances are preferred, since we will have more confidence that our single sample estimate will be as close as possible to the value of the true population parameter (assuming the estimator is unbiased)
- This will mean the we can conduct statistical tests of the significance of regression coefficients accurately, and with the best chance of rejecting the null hypothesis of no effect of X on Y, given a true relationship in the population (i.e., "standard errors" will be accurate and as small as possible)

Desirable Properties of Estimators: Consistency



- Some estimators have desirable properties only in large samples, or asymptotically as N→∞
- Estimators are "consistent" if, as N gets larger and larger, the estimates they produce converge on the true population parameter
- Some estimators are biased in small samples but the bias disappears as N→∞

So the ideal regression estimator would produce a sampling distribution that is centered around the true population slope, would have minimum variance, and would converge on the true population slope as sample size increases. If the following assumptions about the population model hold, OLS (assuringly) has these qualities!! If they do not hold, we need to take corrective action – change our model or change our estimator (or both!!)

Assumption 1: No Specification Error

- This says that the bivariate model you have specified between X and Y is actually the "true" model in the population. If not, then estimating the effect of X on Y in our sample will possibly not yield correct estimates of true β, and will not yield correct estimates of the sample-to-sample variance in β, thus preventing accurate testing of statistical significance. This is the most important assumption of all!!!!
- This assumption is violated whenever:
  - You have omitted relevant explanatory variables
  - You have included irrelevant variables
  - The relationship between X and Y is not linear, or not additive (in multiple regression with several X)
  - There is reciprocal causality between X and Y

- Let's say the true population model is:
- But you assume that the true model is:
- That means that

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$Y_i' = \beta_0' + \beta_1' X_{1i} + \varepsilon_i'$$

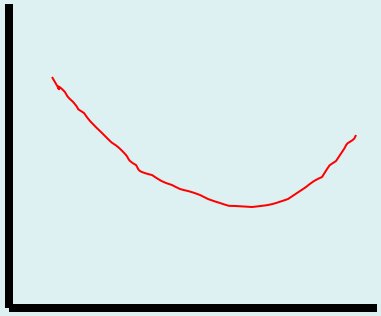
$$\varepsilon_i' = \varepsilon_i + \beta_2' X_{2i}$$

the error term in your assumed model is composed of the "true" error term plus the excluded variable. This is specification error!

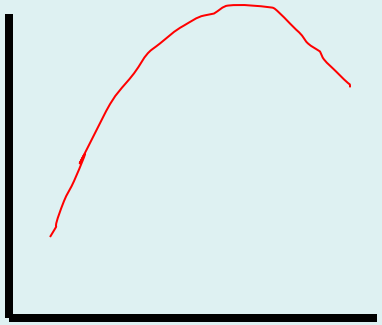
- Consequences for OLS estimates of "β_1":
  - If X_2 is *unrelated* to X_1, OLS b (which is actually an estimate of β'_1) will be consistent but inefficient: since ε' is too big, the estimated variance in b will be too big as well (and R-squared will be smaller than it should be). So harder to achieve statistical significance.
  - If X_2 is *related* to X_1, OLS b will be biased (and inconsistent) as well. The error term ε' will be related to X_1, and some of the true effect of X_2 on Y will be improperly attributed to X_1. Huge problem! (OLS is greedy and maximizes explained sums of squares, even if some actually belong to the error SS)
  - **The technical violation of OLS assumptions is that E(X ε)≠0 (see assumption 4b below)**. Since OLS by construction produces a line where X and ε are unrelated, it is a biased and inconsistent estimator whenever X and ε truly are related in the population. This is one form of **"endogeneity"** in the model

- Other kinds of misspecification
  - Including irrelevant variables in your assumed population model
    - Since they are not relevant for explaining Y, OLS estimates of the effects of the other relevant variables that are included will be consistent
    - They will show up as statistically insignificant (in the long run, at least), so generally have minimal consequences. However, in small samples there could be *some* relationship between the "irrelevant" variable, Y and the relevant X variables that will lead to biased estimates. (As N increases, the "true irrelevancy" of the variable will be revealed).
    - But: it will add another variable to the set of *k* independent variables, and *k* figures in the calculation of "adjusted R-squared" and some other statistics relevant to multiple regression that we will discuss next time. Again, in small samples especially this could cause some problems for your analysis, but generally we don't worry as much about this as omitting relevant variables or other kinds of misspecification

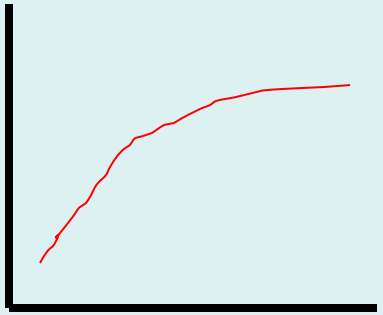
– Functional form is not linear in the population. Examples:



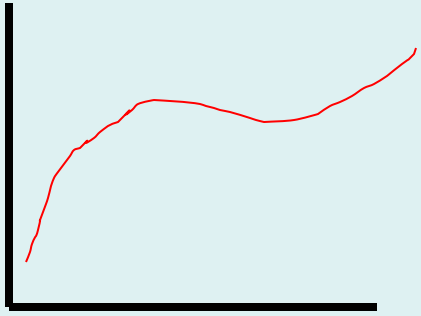
Quadratic "U-curve"



Quadratic "Inverted U-curve"



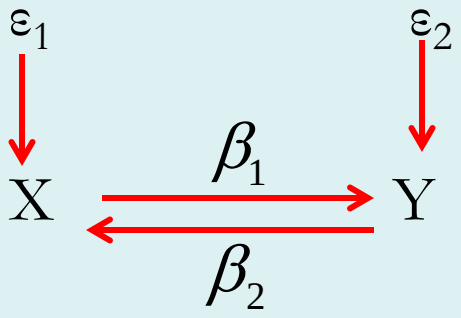
"Logarithmic"



"polynomial" or "cubic"

Estimating the simple linear regression model yields bias. But all of these specifications can be turned into a form that is amenable to OLS linear regression estimation ("linear in the parameters". Either add squared, cube terms to the model or transform the variables.

- Reciprocal Causality as Specification Error
- If true population is a system that looks like this:



then you cannot estimate β_1 or β_2 using OLS. Why not?

You can see, e.g., that ε_2, the error term in Y, is related to X, as ε_2 effects Y which then effects X. So $E(X\varepsilon)\neq 0$ again!

- Using OLS will again attribute some of the effect of Y on X (β_2) as being included in the X on Y effect (β_1). Bias and Inconsistency!!
- This is another kind of "endogeneity" in the X—Y relationship

- **Assumption 2: No Measurement Error in X or Y**
  - Different consequences for OLS if Y or X contains random error
  - In Y:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{and } y_i^* = Y_i + u_i$$

$$\text{then } y_i^* - u_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{and } y_i^* = \beta_0 + \beta_1 X_i + (\varepsilon_i + u_i)$$

  - so OLS overestimates error variance, lower R-squared and larger standard errors!
  - In X:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{and } x_i^* = X_i + v_i$$

$$\text{then } Y_i = \beta_0 + \beta_1 (x_i^* - v_i) + \varepsilon_i$$

$$\text{and } Y_i = \beta_0 + \beta_1 x_i^* + (\varepsilon_i - \beta_1 v_i)$$

  - So OLS overestimates error variance *and* $E(X\varepsilon) \neq 0$ again (since v and x^* are related). So OLS produces biased, inconsistent and inefficient estimates!!!

- **Assumption 3a: There is some variance in X in the population**. This is an "obvious" assumption – if X is fixed at one value, then there are no "changes in X" with which to explain "changes in Y". The OLS procedure breaks down (as the variance in X is in the denominator)
- **Assumption 3b: There is no perfect correlation between the Xs** in a multiple regression population model. This would result in *perfect multicollinearity* between explanatory variables. Intuitively, we could not be able to distinguish the effect of X_1 from the effect of X_2 using OLS (or any other estimator). As we will see in the multiple regression section, estimation is mathematically impossible in the extreme case. As the correlation approaches 1, results become highly unstable.

Assumptions about the Population Model's Error Term (ε)

- **Assumption 4a: The expected value of ε is 0, or $E(\varepsilon_i)=0$**
- **Assumption 4b: The expected covariance of X and ε is 0, or $E(X_i\varepsilon_i)=0$**

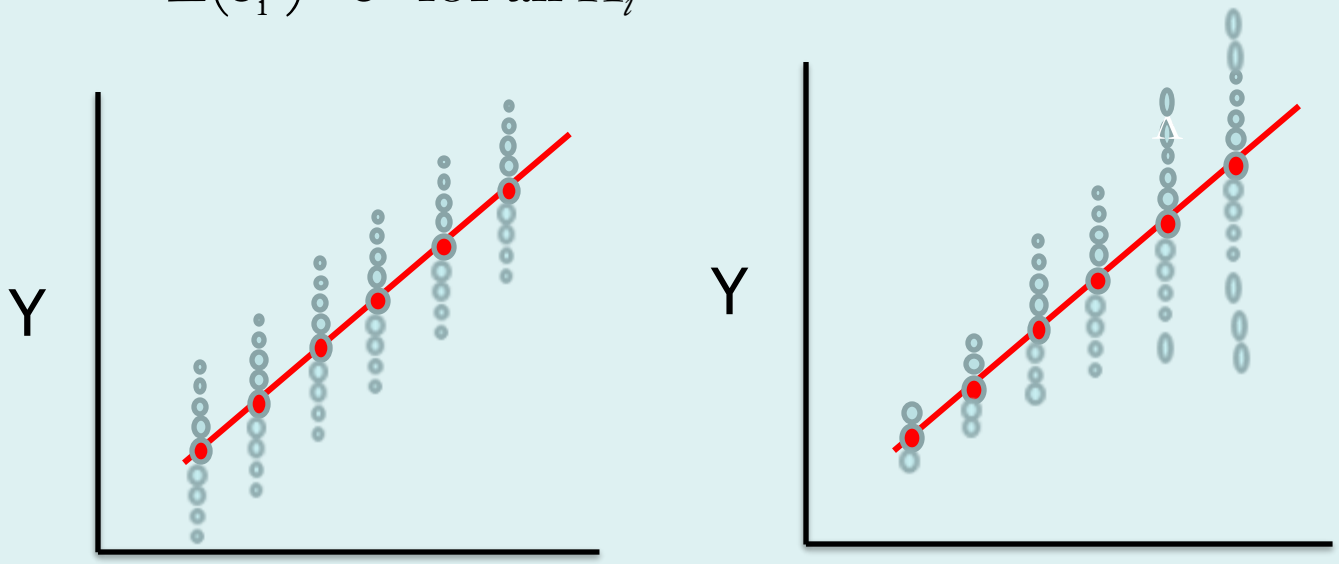
These assumptions summarize the specification error assumption we covered in 1, so there is not much new here aside from the more technical language.

- Since ε_i is all of the omitted variables influencing Y as well as random measurement error, we assume that their average value is neither positive (pushing the conditional mean of Y upward) nor negative (pushing the mean of Y downward)
- Violating this assumption (4a) leads to bias in the estimation of the intercept (though this is usually not crucial to political science inquiry)
- We also assume (4b) that X is independent of the error term (i.e., no endogeneity) for the reasons stated above. Violated when:
  - **The population model contains relevant variables that are related to X which have been omitted**
  - **There is random measurement error in X**
  - **There is reciprocal causality between X and Y**

Assumptions about the Population Model's Error Term (ε)

- **Assumption 4c: Homoskedasticity, or equal error variance at all levels of** X

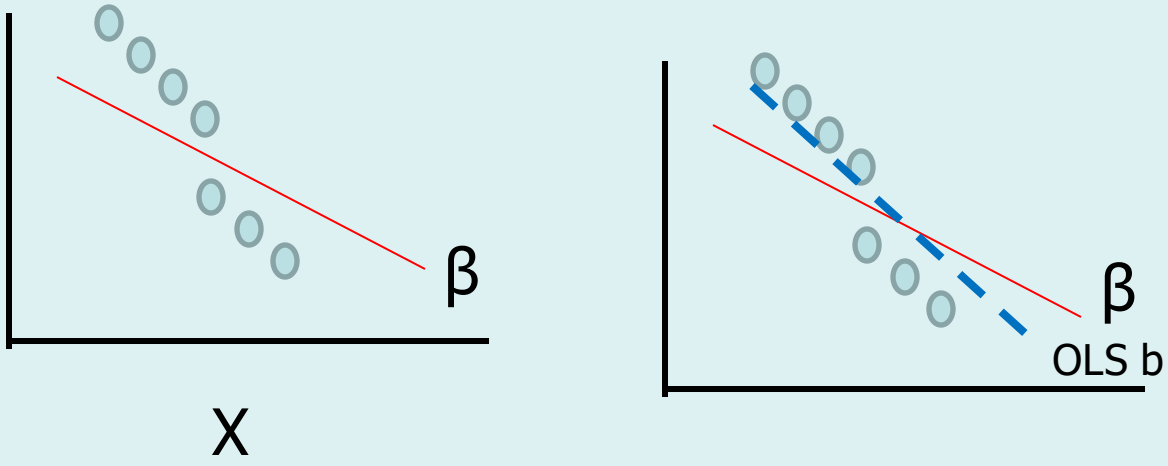
$E(\sigma_i^2)=\sigma^2$ for all X_i



- Can arise because theoretical relationship *is* heteroskedastic, from omitted variables, clustering in data, and other reasons
- Consequences for OLS: unbiased but inefficient – we can use other estimators that produce less variance in the sampling distribution. Weighted Least Squares (WLS), or, at minimum, OLS with "robust" standard errors

Assumptions about the Population Model's Error Term (ε)

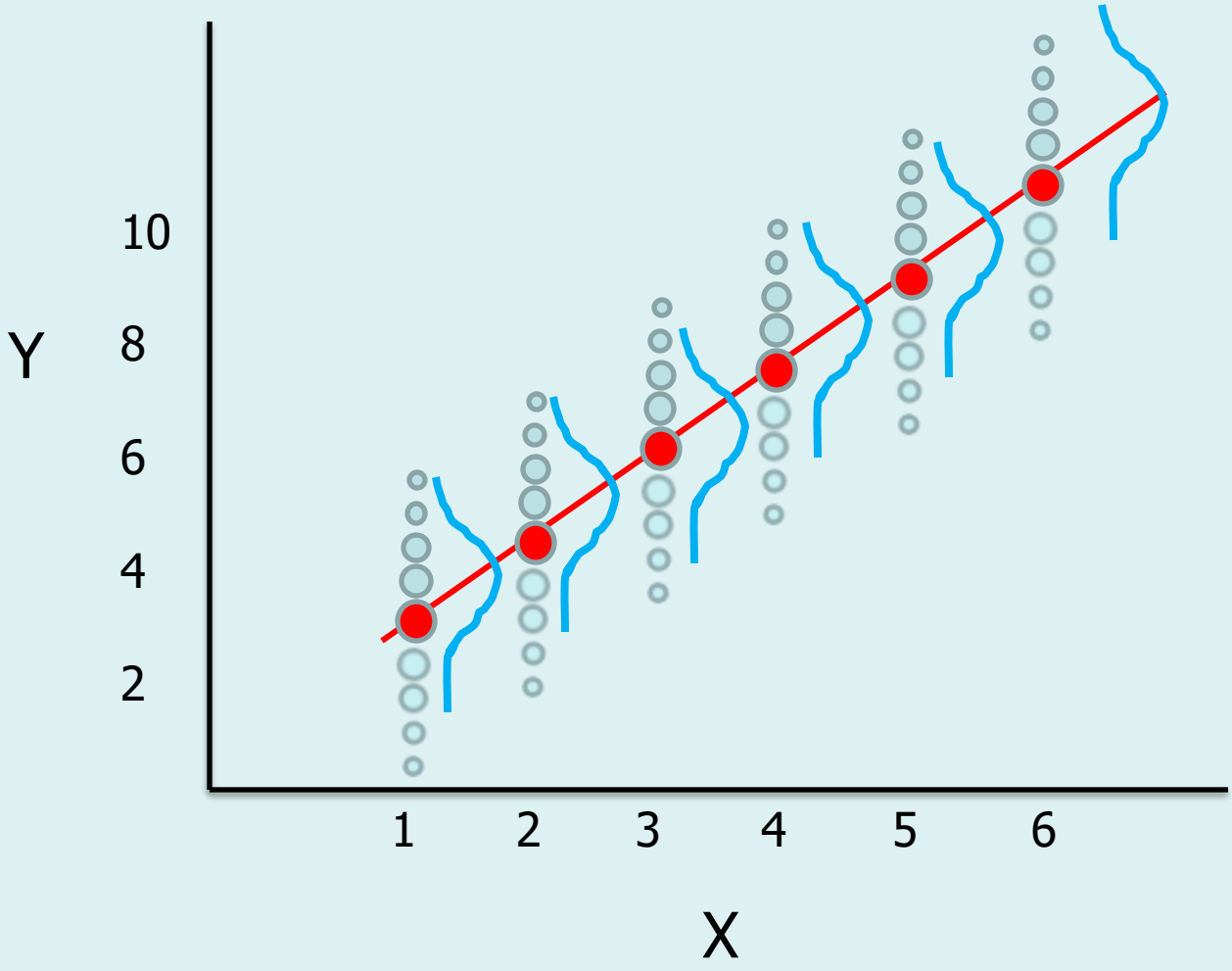
- **Assumption 4d: No Autocorrelation, or $E(\varepsilon_i \varepsilon_j = 0)$ for all *i* and *j***
- Residuals for one case should be unrelated to residual for all other cases. OLS treats all cases equally and wouldn't be able to recognize a situation where the errors are related – it would minimize the sum of squared residuals even though it shouldn't! This problem occurs most often in time-series data.



- Causes: Shocks that persist, omitted relatively stable independent variables
- OLS fits the best line it can, doesn't recognize that Y is where it is at a given point because of past values of the residual. Consequences: Inefficiency and possible bias! Corrections are available ("Generalized Least Squares" or robust standard errors for time-series situations)

- **Assumption 4e Residuals are normally distributed, i.e., $\varepsilon_i \sim N$**
- Above assumptions 1-4d are all that are needed to estimate the population parameters without bias and with minimal variance in the sampling distribution. We can say, following the famous Gauss-Markov Theorem, that the OLS estimator is **BLUE** – the "best linear unbiased" estimator of the population parameters ("best" meaning "minimum variance" here).
- But we still cannot conduct statistical tests because we do not know the exact shape of the sampling distribution, so we cannot rely on statistical theory to tell us, e.g., how far away from a hypothesized population value our sample value is likely to be.
- If we are willing to make the further assumption that the residuals in the population are distributed **normally**, then this problem can be solved
- If $\varepsilon_i \sim N$, then $\beta \sim N$, i.e., the sampling distribution of the slopes will be distributed normally
  - If $\varepsilon_i \sim N$, then Y is distributed normally
  - β (and b) is a linear combination of Y, specifically $b = \frac{\Sigma(X - \bar{X})}{\Sigma(X - \bar{X})^2} Y_i$
  - Any coefficient that is a linear combination of a normal random variable is itself a normal random variable

The Population Regression Function (PRF) with the assumption of normal residuals



Normality (continued)

- Is the assumption reasonable?
  - Yes, if we assume that the residuals are made up of all of the omitted variables in the population model. The Central Limit Theorem (CLT) says that the sum of a large number of independent random variables tends to a normal distribution as the number of variables increases.
  - In small samples, the normality assumption is more important than in large samples. As sample size increases, deviations from normality can be "tolerated" more and OLS will still produce desirable estimates (as the sampling distribution from even abnormally distributed residuals will be unbiased, efficient, and approximately normal (provided the other assumptions hold).
- So, we summarize the error term assumptions as: $\varepsilon_i \sim N(0, \sigma^2)$
- Or, the residuals are normally distributed with a mean of 0, and a variance σ^2 that is constant for all X_i. We can also say that the residuals are "independent and identically distributed", and this subsumes the "no-autocorrelation" idea as well

Sampling Distribution of β Under OLS Assumptions

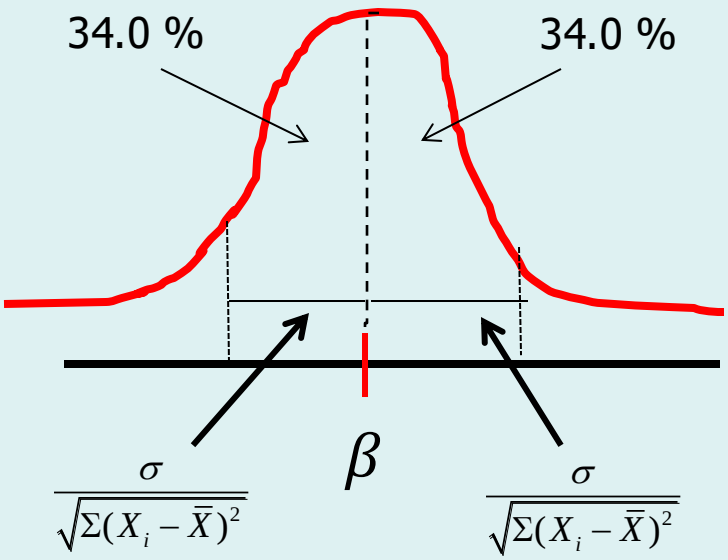
- If all OLS assumptions hold, then it can be shown that

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2})$$

- or, the sampling distribution of $\hat{\beta}$ will be:
  - normal
  - centered around the true population parameter β
  - with a standard deviation ("standard error") of

$$\sigma_\beta = \sqrt{\frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}}, \text{ or } \frac{\sigma}{\sqrt{\Sigma(X_i - \bar{X})^2}}$$

  - the numerator is the population error standard deviation (RMSE), the denominator is the square root of the sums of squares in X
  - this puts us in the position to conduct statistical tests of significance



- Sampling Distribution of β
- OLS is "BLUE":
  - Unbiased, as $E(\hat{\beta})=\beta$
  - Efficient, with minimum variance among class of linear unbiased estimators (i.e., smallest possible standard deviation or "standard error");

  AND
  - the sampling distribution from OLS is normal

3. Hypothesis Testing in Regression Analyses

- As with all statistical analyses, once we have established the nature of the regression relationship between X and Y in our sample, we need to see whether we can reject the idea that there is truly no relationship between X and Y in the overall population. In other words, we need to conduct *statistical inferences* from our sample to the population
- The tool for this in regression is the *sampling distribution of the slope*, shown on the previous slide
- We assume a "null hypothesis", where the true population value of the slope is 0, and we determine how many standard deviations or standard errors away from 0 is our sample value b or $\hat{\beta}$. If it is sufficiently unlikely to have obtained our sample value from a population where the value of β were 0, we reject the null hypothesis (with some probability α of being wrong). We then say that there is a "statistically significant slope coefficient" or "statistically significant bivariate relationship" between X and Y

Steps in Hypothesis Testing

- Specify the null and alternative (or research) hypotheses
  - H_0: $\beta_1 = 0$
  - H_r: $\beta_1 \neq 0$ ("two-tailed"), or $\beta_1 > 0$ ("one-tailed")
- Specify the alpha, or significance level
  - $\alpha = .05$ (the .05 significance level). This means that we will reject the null if the chances of observing, through random sampling, a value of the slope as large as we observed in our sample *if the null hypothesis were true* is less then .05, or 5%.
- Construct a test statistic and make a decision
  - In this case, we have a sampling distribution that is normal, so we can calculate how many *standard errors* away from 0 our sample value is. This would be a z-score, calculated as

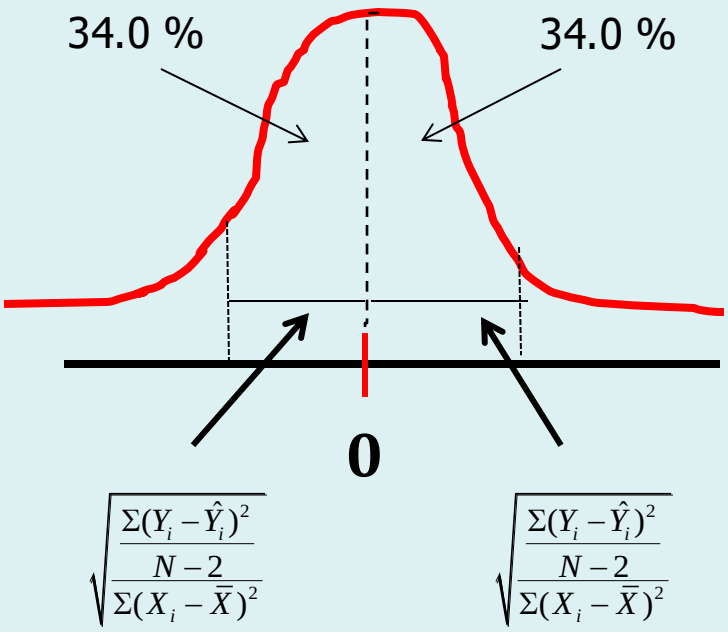
$$z = \frac{(b - \beta_{1_0})}{\sigma_\beta} \text{ or } \frac{(b - \beta_{1_0})}{\frac{\sigma}{\sqrt{\Sigma(X_i - \bar{X})^2}}}$$

- Slight complication, however, in that we do not observe the population residual sum of squares, which is the (squared) numerator in the formula of the standard error of the slope. We need to estimate it from our sample residuals.
- It can be shown that the quantity $$\hat{\sigma}^2 = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{N-2}$$

is an "unbiased" estimate of the population error variance σ^2. So we can substitute that into our calculation of how many "estimated standard errors" our sample slope is from the population value specified in the null hypothesis.

- By using the "estimated" value of σ, we need to use "t" as our test statistic, not "z". "t" is $$t = \frac{(b-\beta_{1_0})}{\hat{\sigma}_\beta} \text{ or } \frac{(b-\beta_{1_0})}{\frac{\hat{\sigma}}{\sqrt{\Sigma(X_i - \bar{X})^2}}} \text{ or } \frac{(b-\beta_{1_0})}{\sqrt{\frac{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{N-2}}{\Sigma(X_i - \bar{X})^2}}}$$

and it is distributed normally as N gets larger (>100 or so). It has a different distribution depending on the ***degrees of freedom*** in the analysis, which is the number of cases that are free to vary, given the constraints in the model. In this case we lose 2 df for the error sums of squares since we needed "a" and "b" to calculate it. So the t-test for the bivariate regression slope has degrees of freedom equal to N-2.



Note: In large samples (N>100), the critical values for t are ±1.96 for the .05 significance level (two-tailed), ±1.65 (one-tailed)

- So where is our slope "b" on this estimated sampling distribution? How many "t" values away from 0, and what were the chances of observing a "t" that large *if the null hypothesis were true*? If less than .05, we reject the null.
- Our week 1 exercise example: b=1.2

$$t = \frac{(b-\beta_{1_0})}{\sqrt{\frac{\Sigma(Y_i-\hat{Y}_i)^2}{\frac{N-2}{\Sigma(X_i-\bar{X})^2}}}} = \frac{(1.2-0)}{\sqrt{\frac{26.84}{\frac{8}{30.4}}}} = \frac{1.2}{.33} = 3.64$$

- With 8 df, we reject the null at the .05 level (two-tailed) if t>2.31 or t<-2.31. These are the **critical values** for t at that df.
- Our decision: Reject H_0! The chance of observing a b of 1.2 or greater if the null hypotheses were true is less than .05. (STATA provides the exact probability as "p>|t|"; you want a value at or below .05)

Notes

- Examine formula for estimated standard error of the slope

$$\hat{\sigma}_{\beta} = \sqrt{\frac{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{N-2}}{\Sigma(X_i - \bar{X})^2}} = \frac{\hat{\sigma}}{\sqrt{\Sigma(X_i - \bar{X})^2}}$$

- The numerator is the RMSE or the Standard Error of Estimate
- The denominator is the square root of the Sum of Squares in X
- This means that, other things being equal, smaller standard errors (and greater chances of rejecting null hypotheses) are the result of:
  - Smaller error sums of squares (or larger R-squared from the model)
  - More variation in X
  - Larger N of cases
- Be careful not to confuse statistical versus substantive significance!
- Pay attention to "low power": not large enough N to detect effects with sufficient confidence, so there is the risk of **not rejecting false nulls ("Type 2 error")**

Interval Estimation for β

- We sometimes want to estimate "confidence intervals" around our estimate for β. Our best guess of the population parameter β is "b", our sample estimate.
- We can construct a 95% confidence interval around "b" according to the formula: $b \pm t_{.025} * \hat{\sigma}_{\beta}$

  where $t_{.025}$ is the critical value for *t* at the .05 significance level for the given number of degrees of freedom
- The critical value at the .05 level (95% confidence) will be 1.96 for N>100. The critical value at the .10 level (90% confidence) will be 1.65 for N>100.
- So take your sample estimate and go out 1.96 standard errors on either side to construct the 95% confidence interval for the population parameter β.
- If β was not in that interval, it would have been very unlikely (less than 5% of the time) to have observed the sample value of "b" that we did observe. Still, it could have happened! **(That is why we only have 95% confidence!)**.
- **Relationship between confidence intervals and hypothesis testing: If the confidence interval overlaps 0, then a t-test will show that you cannot reject the null hypothesis**

Testing the Significance of the Entire Equation

- Can also conduct a general test of the overall significance of the equation in terms of explaining variation in Y
- In multiple regression, this would test the null hypothesis that *all* slopes in the population associated with all explanatory variables are 0:

$$H_0: \beta_1 = \beta_2 = \beta_j = 0$$

- This is in effect a test of the statistical significance of R-squared; does the equation as a whole explain a significant amount of regression sums of squares, or could the R-squared that we obtained in our model have come about through random sampling error?
- The test statistic here is **F**, and it is calculated as the ratio of Explained to Unexplained Variation in Y, or:

$$F = \frac{\text{Regression Variance}}{\text{Error Variance}} = \frac{\dfrac{\text{Regression Sums of Squares}}{df\,(regression)}}{\dfrac{\text{Error Sums of Squares}}{df\,(error)}} = \frac{\dfrac{\Sigma(\hat{Y}_i - \bar{Y})^2}{k}}{\dfrac{\Sigma(Y_i - \hat{Y}_i)^2}{N-k-1}}$$

- We compare the value of F with the critical value at a given significance level, given the df in the numerator and the denominator, and make a decision about H_0

- In Analysis of Variance language, F is the ratio of two "Mean Squares"
- A Mean Square (or a "Variance") is a Sum of Squares divided by its associated degrees of freedom. Here:

$$\text{Mean Square Error ("Residual")} = \hat{\sigma}^2 = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{N - k - 1}$$

$$\text{Mean Square Regression ("Model")} = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{k}$$

- The degrees of freedom associated with the error sums of squares is N-k-1, where k is the number of independent variables. In bivariate regression, this is N-2 as we saw (we lose two df in calculating the intercept and the one slope coefficient).
- The degrees of freedom associated with the regression sums of squares is *k*.
- The total degrees of freedom in a sample is N-1 (we lose 1 df in calculating the sample mean).
- The quantities here correspond exactly to those on the STATA output, where "Mean Square Regression"="Model" and "Mean Square Error"="Residual"

- It can be shown that

$$\text{E(Mean Square Regression)} = \sigma^2 + \beta^2 \Sigma (X_i - \bar{X})^2$$

which is equal to σ^2 under the null hypothesis that $\beta=0$.

So under the null, we would expect an F of 1, since both the numerator and denominator are separate estimates of the same quantity (σ^2).

- Can find the critical value for F with df (numerator) and df(denominator) and compare the obtained F to this value. If exceeds, reject H_0. If not, do not reject.
- In bivariate case, $F=t^2$ (so always come to same conclusion regarding H_0)

- Logic of F: does the model have a significantly greater amount of **regression sums of squares**, divided by its df, than it **does error sums of squares**, given its df, than we would have expected by chance, or through random sampling error?
- Can see the relationship of F to R-squared:
- $R^2 = \frac{Regression\ SS}{Total\ SS} = 1 - \frac{Error\ SS}{Total\ SS}$
- $F = \frac{\frac{Regression\ SS}{k}}{\frac{Error\ SS}{n-k-1}}$ then divide numerator and denominator by TSS
- $F = \frac{\frac{Regression\ SS}{Total\ SS * k}}{\frac{Error\ SS}{Total\ SS * (n-k-1)}} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$
- So F is in effect a test of significance of R^2!

Statistical Inference in Stata Output

```
. regress var2 var1

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =   13.14
       Model |  44.0644737     1  44.0644737            Prob > F      =  0.0067
    Residual |  26.8355263     8  3.35444079            R-squared     =  0.6215
-------------+------------------------------           Adj R-squared =  0.5742
       Total |      70.9       9  7.87777778            Root MSE      =  1.8315

------------------------------------------------------------------------------
        var2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        var1 |    1.203947   .3321798     3.62   0.007     .4379393    1.969955
       _cons |    2.006579   1.269257     1.58   0.153    -.9203338    4.933492
------------------------------------------------------------------------------
```

Standard error of b: $\hat{\sigma}_b$

t

Probability of obtaining a "t" of 3.62 or greater if the null hypothesis were true:

$$F = \frac{\frac{Regression\ SS}{k}}{\frac{Error\ SS}{n-k-1}}$$