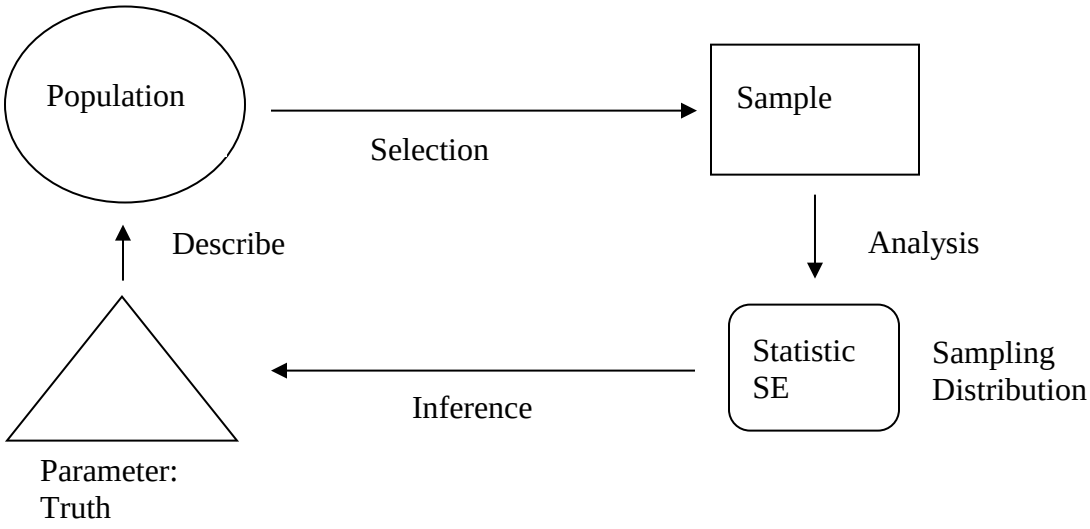
A Research Paradigm: A Model for Empirical Research



There is a basic research paradigm that we use when engaging in statistical analysis. It generally focuses on our ability to describe important phenomena in the population. Sometimes those phenomena are relatively simple, like level of education or support for a specific policy. Most of the time the phenomena in which we are interested are more complex, such as the association between socio-economic status and achievement or the impact of an intervention on closing the achievement gap. Often they are even more complex, or multivariate, such as the relation between teacher quality and student achievement in mathematics, reading, writing, and science.

Whatever the level of complexity, we generally cannot study the entire population. So we select a sample, do analysis, estimate population parameters with statistics and associated estimates of precision or sampling error, from which we make an inference about the parameter in hopes of being able to describe the phenomena of interest in the population.

This cycle of sampling, statistical estimation, and inference about a population requires a systematic treatment of scientific methods and rigor – more complex phenomena and more complex sampling require more rigor and careful research design.

The Basis of Inferential Statistics

For most empirical research, we rely on sampling distribution theory and the Central Limit Theorem to support our ability to make inferences about the population.

These are based on the known properties of the Normal Distribution $\sim N(\mu, \sigma^2)$

Consider the estimate of the mean: $\overline{X} = \frac{\sum X_i}{N}$

From sampling distribution theory, we find that the sample mean has a distribution:

$$\overline{X}_j \sim N\left(\mu, \frac{\sigma^2}{N}\right) \qquad SE\left(\overline{X}_j\right) = \sqrt{\frac{\sigma^2}{N}}$$

Parameter: θ
Sample: N
Estimator: $\hat{\theta}$ $\qquad SE\left(\hat{\theta}\right) = \sigma_{\hat{\theta}} \qquad$ CI for θ: $\hat{\theta} \pm c_{\alpha/2}\sigma_{\hat{\theta}}$

Above, we have the basic structure for inferential statistics. We rely heavily on the properties of the central limit theorem that allows us to make inferences from statistics to populations, but this requires minimum sample sizes. In fact, we know, from the central limit theorem, that the larger the sample (to some extent), the more precise our estimates – the closer a given sample estimate will be to the population parameter.

We can estimate a statistic, like the mean. Sampling distribution theory demonstrates that as we take all possible samples of size *N*, we can observe the distribution of sample means – what are all the possible values of a mean that can be obtained from a sample of *N*. We rely on the Central Limit Theorem to know that the distribution of sample means has a mean itself – which is the population mean, and sample means have a variance. The standard deviation of the sample means is the standard error of the mean – the standard deviation of means of size *N* drawn from a population. This provides an indication of the variability in means we might observe from different samples – or – the typical error in any given mean as a sample of the population.

With this information, and because the Central Limit Theorem tells us that with samples of at least 30, the sampling distribution of the mean is normal, no matter what the shape of the population distribution, we can make inferential statements about the population mean. We can use the standard error to estimate confidence intervals, such that a known proportion of such confidence intervals will contain the population mean – a function of the standard normal curve.

A statistical paradigm

Whenever we engage a statistical model to make inferences about the population, we must engage in three steps:

1. Model Building
   Specify the model for your situation; the most important step

2. Estimation of Parameters
   Getting results; the third most important step

3. Testing Fit of the Model
   Consistency between the data and the model; the second most important step

Model specification is clearly the most important step and should receive the most effort and attention. Model misspecification is likely the most serious deficit facing educational and social-science researchers today.

Testing the fit of the model to the data is the second most important step, but far too frequently, ignored completely. If the model-data fit is poor, there is no way to support interpretation of the estimated parameters – no matter how significant they appear.

Once we do these things well, we can interpret the estimated parameters.

Each of these three phases is described next.

Building Linear Models for Data Analysis

1. Theory

   Always begin with theory. Develop an argument, supported by previous literature (could combine several different sources) and add a personal touch.

2. Model specification (outcome[s], explanatory variables)

   Define all factors/variables involved in the theory. Draw a diagram of the relationships among the variables. Specify the outcome(s), explanatory variables, mediating/moderating variables, potentially confounding variables. Argue causality based on the design; beware of the term "predictor."

3. Measuring Variables (reliability and validity)

   Using standardized instruments versus self-constructed instruments. Standardization population should be recent, representative, and relevant. Self-constructed instruments must be piloted and evaluated. Most direct measurement possible is best.

4. Data collection: sampling (random, convenient, purposive)

   Affects some statistical manipulations; most assume samples are randomly drawn from an identifiable population. A given statistic may not be dependent on sampling method, but the inference is always dependent on sampling method and research design.

Estimation of Parameters

1. Factors in the model can be fixed or random

   Fixed factors are variables in which the data in your sample represent all possible levels (scores, groups, treatments, behaviors, conditions) in the population to which you generalize.

   Random factors are variables in which the data in your sample represent a subset of levels from the population (which is infinite) sampled with a known model – and you wish to generalize to the entire population of all possible levels.

2. General Linear Model Assumptions

   a. Structural assumptions allow us to interpret the results
      i. Observations are independent
      ii. Variables are linearly related
      iii. Explanatory variables are independent
      iv. Explanatory variables are measured without measurement error
      v. The right variables are in the model (argumentation: confounds, misspecification)

Regarding Structural assumptions, these are the assumptions that allow us to interpret results. The assumption of independent observations is related to the stochastic assumption regarding independence of residuals – they coincide. This allows us to consider each observation as a unique independent contributing piece of information. To the extent that observations provide common or related (dependent) information, as in the example of students nested in classrooms such that their academic self-efficacy is all a function of their teacher's instructional style, we violate local independence of observations and should model the within-classroom dependency – perhaps using a multilevel model. Dependency across observations will

Variables being linearly related is a simple function of the linear model – we are only estimating the linear component of associations among variables. So if the variables are nonlinearly related, we underestimate their association.

Explanatory variables are independent – helping us avoid the problems with multicolinearity. To the extent that explanatory variables are dependent, we are unable to partition variance among the variables – they are confounded.

Most linear models assume that explanatory variables are measured without measurement error – considering all variance to be true variance. There is no way to directly partition out error variance, unless a measurement model accompanies the estimation, such as in a structural equation model where we estimate latent variables and then associations among the latent variables – which are partitioned from the measurement error in the indicator variables. Measurement error results in underestimating associations and effects.

Finally, the right variables are in the model – that is, the model is correctly specified. This is the source of most research efforts in education and the social sciences. We believe we can do a better job of estimating effects than the previous researcher, because we have a better set of variables to account for the phenomena of interest.

b. Stochastic assumptions allow us to test parameter estimates
    i. Errors have a mean of zero
    ii. Errors have constant variance; homoscedasticity
    iii. Errors are normally distributed
    iv. Errors are independent
    v. Errors are independent of explanatory variables

Regarding the stochastic assumptions, these are the assumptions that allow us to test parameter estimates regarding their statistical significance – the extent to which they are likely to exist in the population. The assumption that errors have a mean of zero is critical because their overall effect on a parameter estimate then is zero – statistical estimates of parameters are unbiased. In OLS, the model is defined to minimize the differences between observed and predicted values, equalizing the errors that are positive and negative, so the mean is zero.

Errors also are assumed to have constant variance or be homoscedastic. This allows us to estimate a single value for the residual or error variance – such as the standard error of the estimate – the standard deviation of residuals about the regression line. Employing a single standard error is key to most statistical tests. We will see this shortly. Heteroscedasticity leads to increased Type I error rates.

Errors are normally distributed allows us to make probabilistic statements about the likely size of an error or residual – based on the standard normal curve (68% of values fall within +/- 1 SD). This is also somewhat intuitive, that most errors are quite small and few errors are extreme. Non-normal error distributions can result from outliers, among other things, and result in problematic standard error estimation and incorrect confidence intervals.

Errors are independent – related to the independence of observations – allows us to decompose or partition variance in a simple manner – total variance can be partitioned into variance due to prediction and variance due to error. The error terms are independent so their variance can be estimated in a direct way – based on squared deviations from the mean (of zero), simply the squared errors. Sometimes this is violated in time-series data, where other error models like auto- or serial-correlation are characteristics.

Errors are independent of explanatory variables, providing us with the freedom to interpret the effect of explanatory variables directly. In the event that these last few assumptions are violated, such that errors are not independent or are related to the explanatory variables, this is often an indicator of model misspecification (potential confounding or omitted variables) – leading to complex associations among estimated parameters, preventing us from testing the statistical significance of our statistics as estimates of the parameters of interest.

Additional assumptions in multivariate contexts:

> If the X variables are thought of as random, then we make the general assumption that the joint distribution of Y and the Xs is multivariate normal. If the X variables are fixed, we assume the conditional distributions of Y (given the Xs) are independently and normally distributed. Moderate departures of either set of assumptions are tolerable.

Testing the Fit of the Model & Related Issues

The methods for testing the fit of the model depend on the model being tested. But, these tests and principles build on each other as models become more complex – the simpler modeling requiring simple tests, more complex models requiring more complex tests. We start with the principle of parsimony. Sometimes, we lose interpretability and meaning by building more complex models – where simple is better. But, simple may also be misleading and not tell the whole story.

1. Parsimony
   The simpler model is better

2. Correlations
   Squared correlations tell you the % of variance explained (coefficient of determination).

3. Simple Regression

   R is the correlation between the outcome and the explanatory variable.

   R^2 is the same as the squared correlation between the outcome and explanatory variable. It is a variance accounted for statistic.

   Remember, $R^2 = \frac{SS_{Regression}}{SS_{Total}}$. To test the hypothesis that $R^2 = 0$, $F = \frac{\hat{R}^2 / k}{(1-\hat{R}^2)/(n-k-1)}$,

   with k and n-k-1 *df*. This is equivalent to $F = \frac{MS_{Regression}}{MS_{Residual}}$.

   Also: check the size of the Standard Error of the Estimate (standard deviation of residuals). We know that as you add predictors to a model, the variance explained will generally increase. However, if those predictors are adding more noise than signal, the standard error of the estimate will increase – indicating less precision in prediction. We can increase explained variance while reducing precision.

4. Multiple Regression

   Here, R is the multiple correlation between the model-predicted scores $\hat{Y}$ and the observed scores Y.

   R^2 is the squared multiple correlation; the percent of variance explained in the outcome by the linear combination of explanatory variables. Standard Error is analogous to the size of the average error of prediction, the standard deviation of the residuals as above.

5. Analysis of Variance

Eta-squared, η^2, is an estimate of the maximum squared correlation between the independent variable and the dependent variable – it can be treated as any squared correlation, the proportion of variation accounted for.

$$\eta^2 = \frac{SS_{between}}{SS_{total}}$$

Eta-squared is generally biased upward when based on sample data. Omega-squared is an adjusted value that is better for most purposes.

In Analysis of Variance, the test of mean differences is one of whether the variation between groups is greater than the variation remaining within groups

$$H_0 = \sum_{j=1}^{J}\left(\mu_j - \mu\right)^2 = 0 \qquad F = \frac{MS_B}{MS_W}$$

6. Controlling overall Type-I error rate (α)

Compute a test-wise α to control the overall study-wise α when conducting multiple tests on the same data: $1-\sqrt[c]{(1-\alpha)} \le \alpha$; where *c* is the number of statistical tests or contrasts conducted, and α is the test-wise Type-I error rate used to determine statistical significance for each statistical test on the same data set.

Statistical Modeling

Observations are composed of true (systematic) & error (random) components: $X_i = T_i + E_i$

The notation of Regression

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

where:

Y_i is the outcome value for individual *i*, where $i = 1, \ldots, N$
$\beta_0 \ldots \beta_P$ are parameters
$X_{1i} \ldots X_{Pi}$ are known constants associated with the *P* explanatory variables
ε_i are iid (independently and identically distributed)

In terms of estimation:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_P X_{Pi} + e_i$$

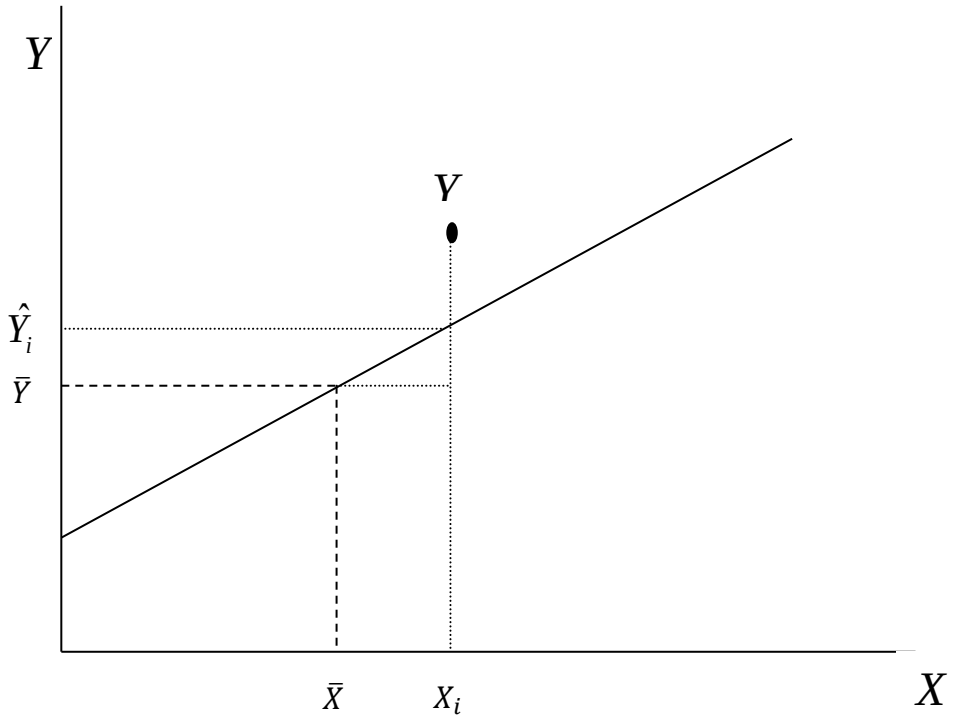
$$Y_i = \hat{Y}_i + e_i$$

This is the simple statement: the observed value equals the sum of the predicted value plus error.

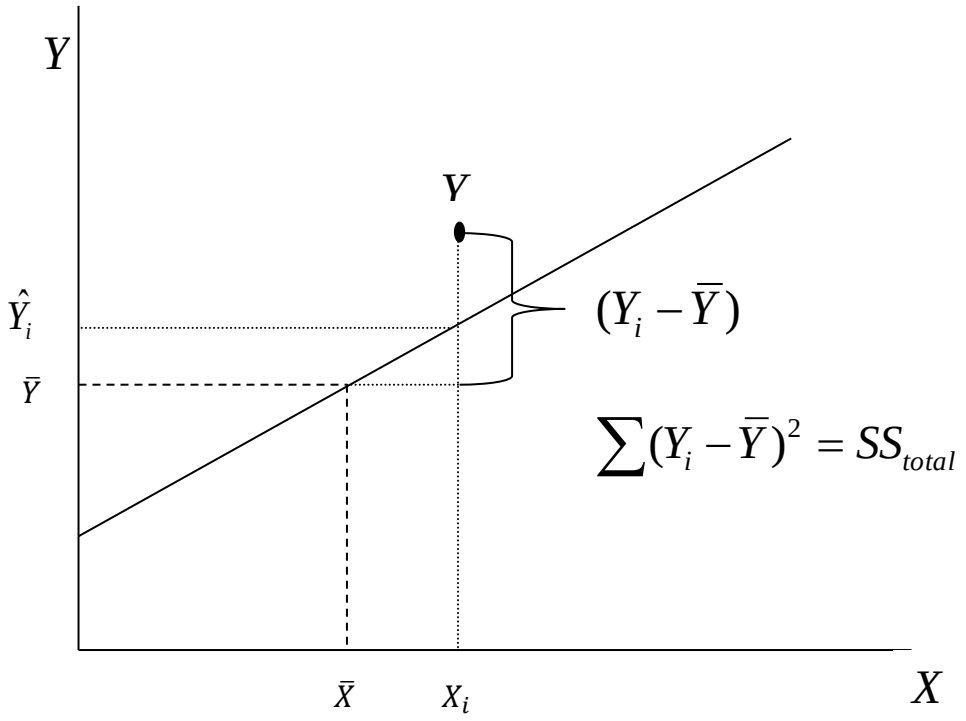
Since the error terms are assumed to be independent of the predictors and predicted values (and homoscedastic), the variance in *Y* is partitioned through sums of squares into two independent components:

$SS_{total} = SS_{regression} + SS_{residual}$

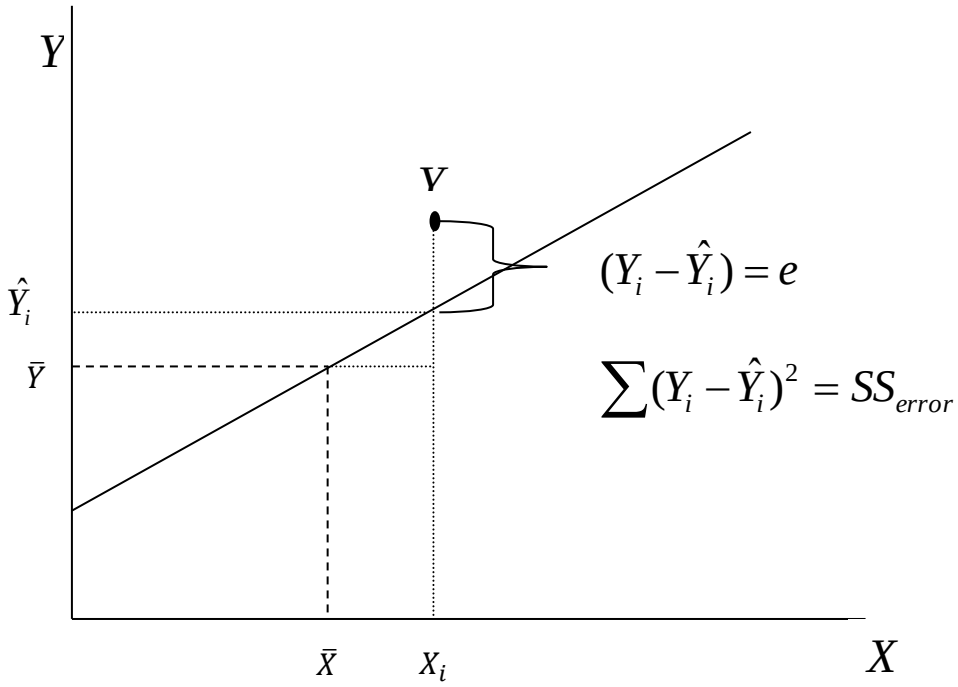
$$\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{N}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$$



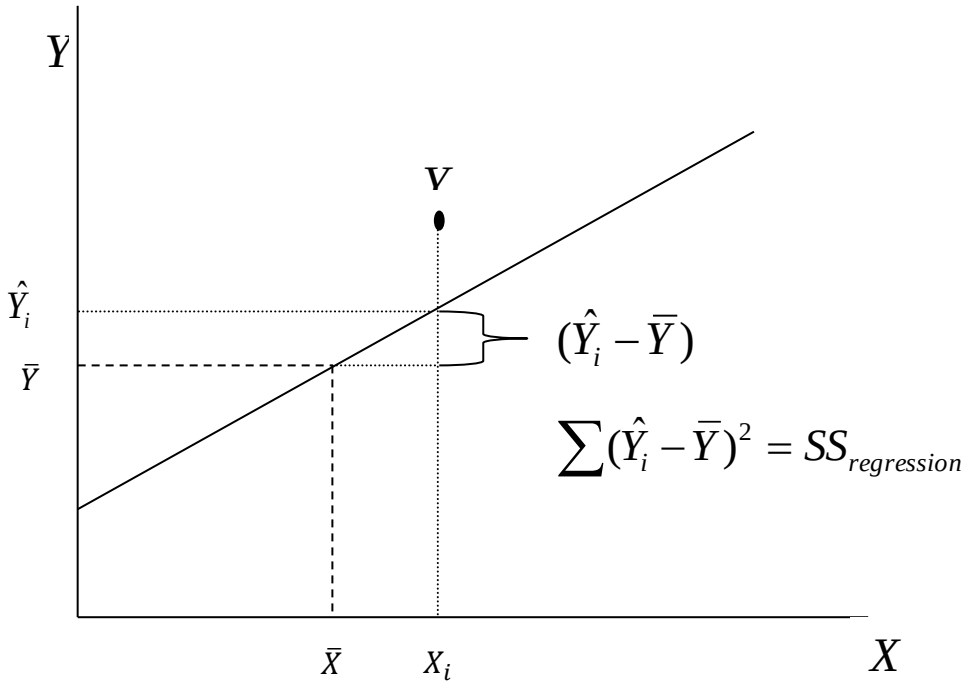
X_i is the observed score on X, and Y_i is the observed score on Y.
The mean of X and Y are marked and intersect on the OLS line.



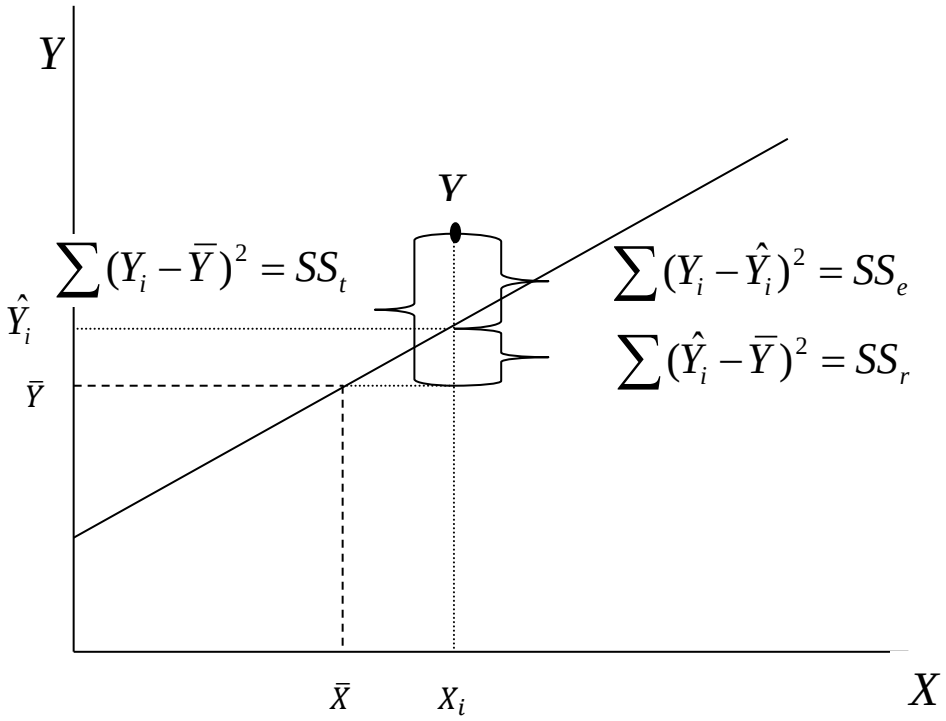
Deviations can be represented as the difference between the observed and mean scores.
Total sums of squares is based on the summed squared deviations of observed values from the mean.



Consider the score of X_i. The OLS line predicts a value of
Even though we know the real value for individual *i* is Y_i .
The difference between the observed value and the predicted value is the residual.

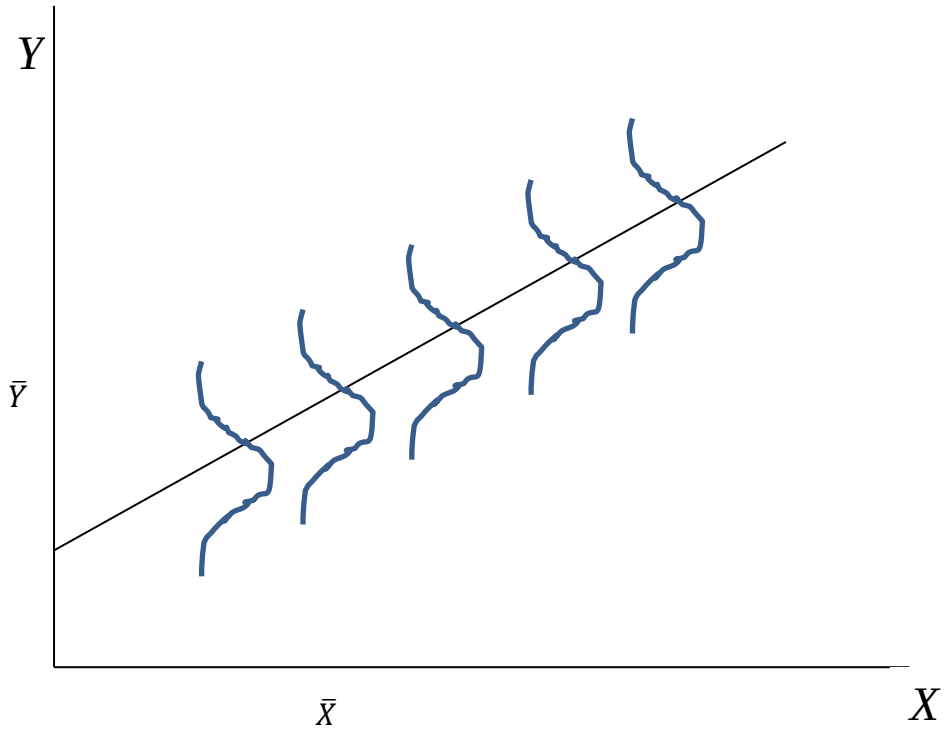


Without information on *X*, our best guess of the associated *Y* value is the mean. The regression improvement of this best guess is the difference between the OLS predicted value and the mean of *Y*.



$SS_{total} = SS_{regression} + SS_{error}$

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$



In regression, we consider conditional distributions of *Y* given *X*. The SD of these distributions of *Y* conditional on *X* is the standard error of the estimate, SE_e. The assumption of homoscedasticity suggests a constant SE_e across values of X.

The notation of Analysis of Variance

In ANOVA, we examine variation in *Y* for individual *i* in group *j* as a function of the group means μ_j and the residual ε for individual *i*. We can also partition this variation as a function of a grand mean μ, group deviations from the grand mean α_j, and individual residuals.

$$Y_{ij} = \mu_j + \varepsilon_{ij} \qquad\qquad Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

where:

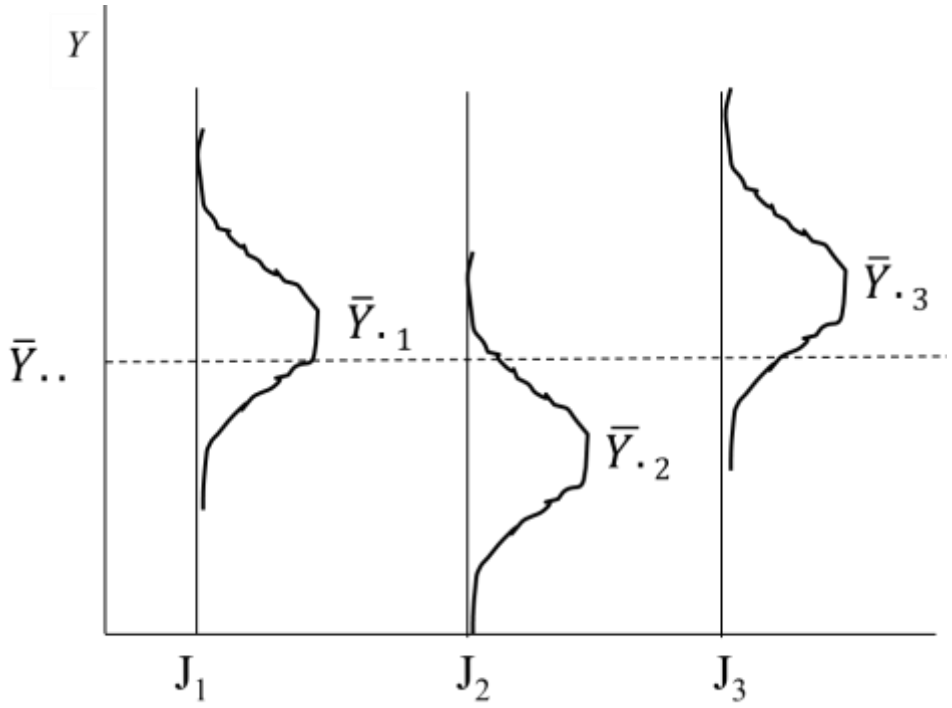
Y_{ij} is the outcome for individual *i* in group *j*, where $i = 1, \ldots, N$; and $j = 1, \ldots, J$.
α_j are parameters, deviations between group means and the grand mean
ε_i are iid (independent and identically distributed)

In ANOVA we condition *Y* on *X* partitioning the distribution by *X*

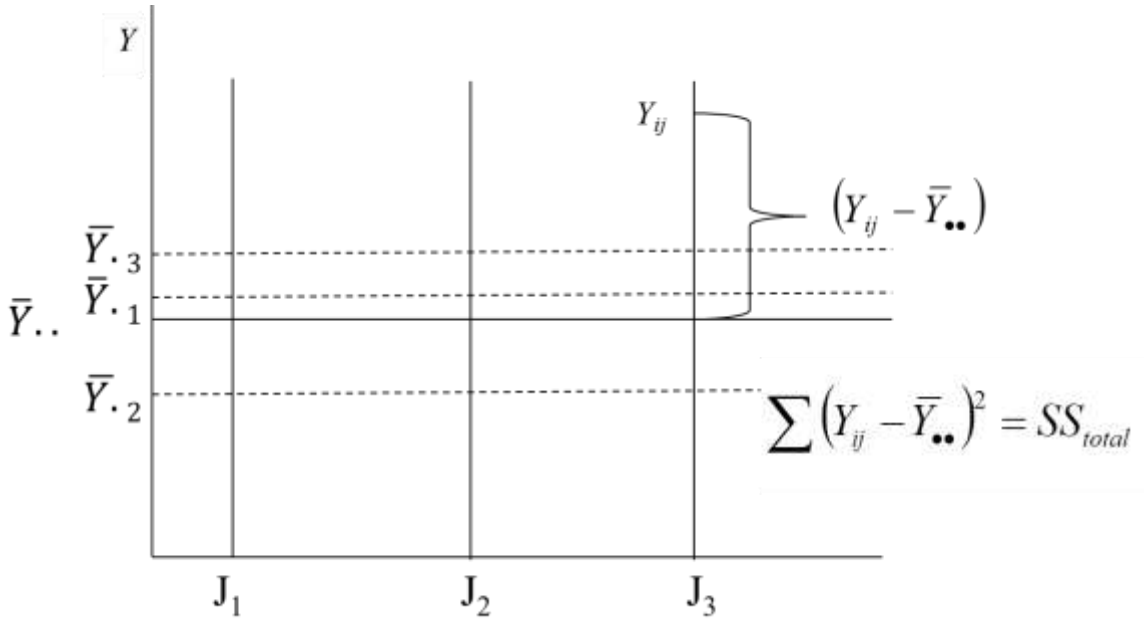
Variance in *Y* is partitioned through sums of squares:

$SS_{total} = SS_{between} + SS_{within}$

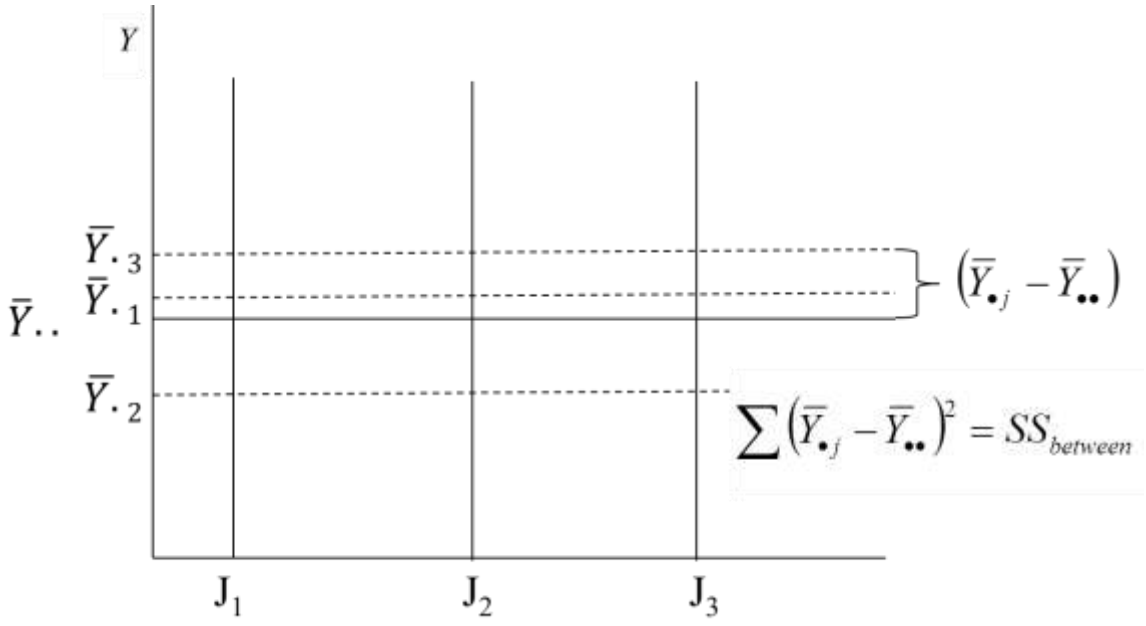
$$\sum_{i=1}^{N}\left(Y_{ij} - \bar{Y}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{N}\left(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}\right)^2 + \sum_{i=1}^{N}\left(Y_{ij} - \bar{Y}_{\bullet j}\right)^2$$



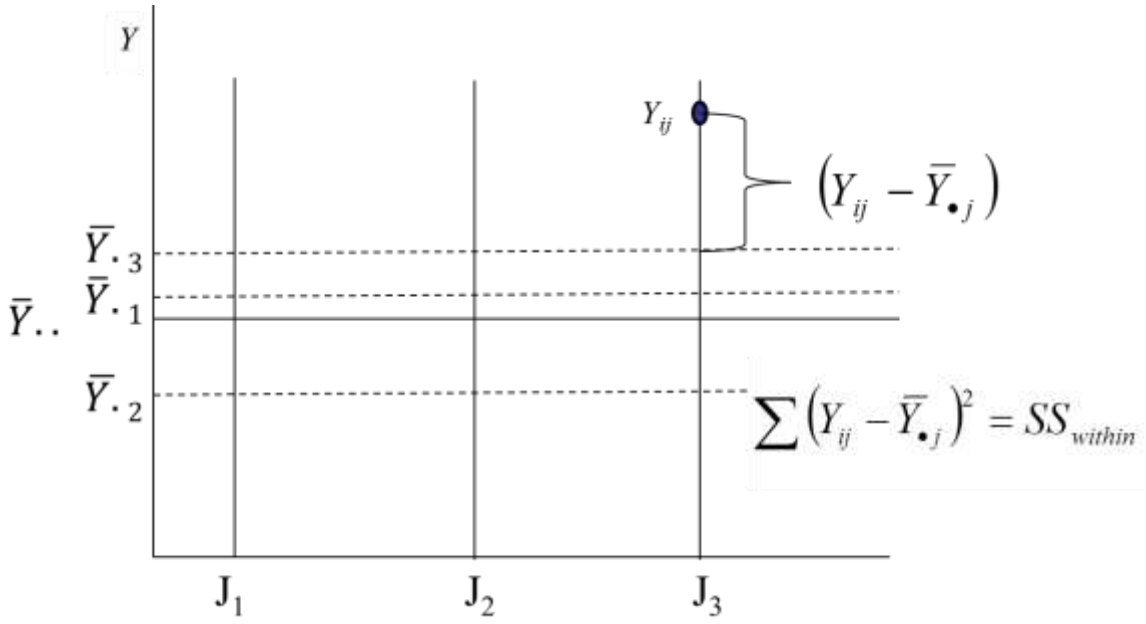
The distribution of scores on *Y* for three groups are illustrated.
Each distribution has a mean and variance. The assumption for ANOVA is homogeneity of variance across groups.



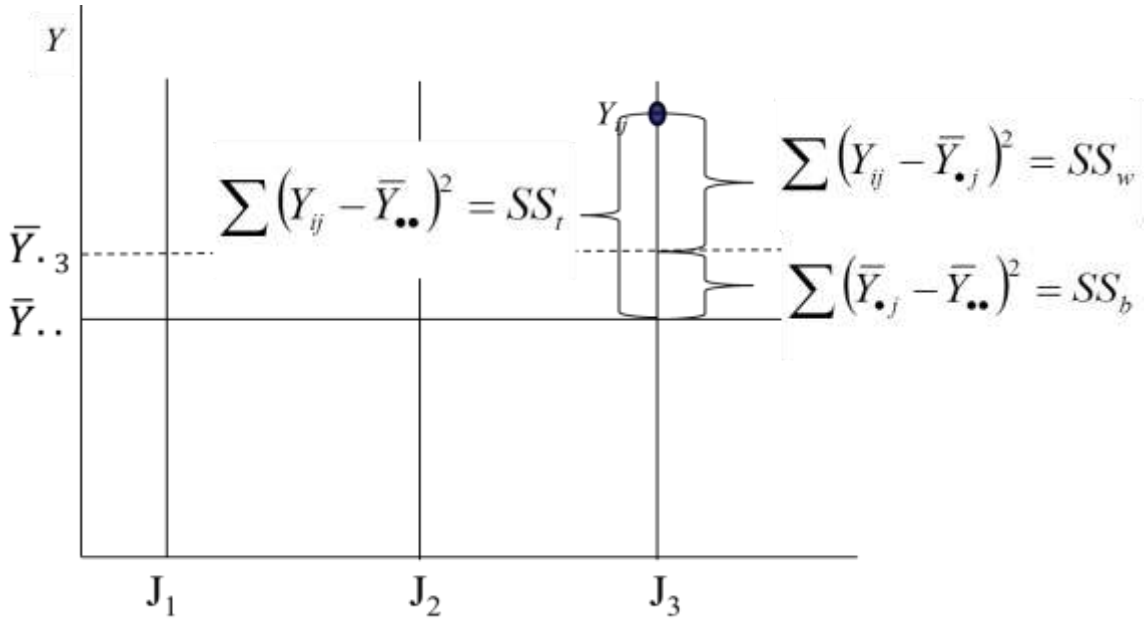
The deviation on *Y* for individual *i* in group *j* is illustrated above.
The total sums of squares is the sum of squared deviations across all groups.



In the absence of information about group members, our best prediction of score on Y is the grand mean. Deviations between group means and the grand means constitute between group scores.



The deviation on *Y* for individual *i* in group *j* is illustrated above.
The total sums of squares is the sum of squared deviations across all groups.



SS_{total} = $SS_{between}$ + SS_{within}

$$\sum\left(Y_{ij}-\bar{Y}_{\bullet\bullet}\right)^2=\sum\left(\bar{Y}_{\bullet j}-\bar{Y}_{\bullet\bullet}\right)^2+\sum\left(Y_{ij}-\bar{Y}_{\bullet j}\right)^2$$

Note that the partitioning of sums of squares for regression and ANOVA is equivalent. This equivalence leads us to the General Linear Model.

Regression:

$SS_{total} = SS_{regression} + SS_{error}$

$$\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{N}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2$$

ANOVA:

$SS_{total} = SS_{between} + SS_{within}$

$$\sum_{i=1}^{N}\left(Y_{ij} - \bar{Y}_{\bullet\bullet}\right)^2 = \sum_{i=1}^{N}\left(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet}\right)^2 + \sum_{i=1}^{N}\left(Y_{ij} - \bar{Y}_{\bullet j}\right)^2$$

$$\underline{\mathbf{y}} = \mathbf{X}\,\underline{\mathbf{b}} + \underline{\mathbf{e}}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

When we employ a design matrix in Analysis of Variance, the model is parallel to the regression model with a matrix of explanatory variables.

A data matrix will contain rows of cases and columns of variables; for example:

ID	SAT	GPA	Gender	IQ
1	560	3.0	1	112
2	780	3.9	0	143
3	620	2.9	0	124
4	600	2.7	1	129

Consider a study where SAT is the outcome variable of interest and the others are explanatory variables.

$$\underline{\mathbf{y}} = \begin{bmatrix} 560 \\ 780 \\ 620 \\ 600 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 3.0 & 1 & 112 \\ 3.9 & 0 & 143 \\ 2.9 & 0 & 124 \\ 2.7 & 1 & 129 \end{bmatrix}$$

Multivariate Regression, ANOVA

Multiple Outcomes, generally correlated
Assumption: multivariate normality

$$\mathbf{Y} = \mathbf{X}\,\mathbf{B} + \mathbf{E}$$

GLM is the general expression of partitioning of variance conditioned on the continuous P variables (*X*s) or conditioned over J groups.