Experimental Design Identification

To correctly design an experiment, or to analyze a designed experiment, you must be able to look at a design situation and correctly assess the salient aspects of the design. I will ask you to identify the design, the treatment variable, dependent variable, degrees of freedom error, experimental unit, sampling unit (if any), and if the treatment is fixed or random.

To begin with, determine what the investigators are trying to do and what they plan to measure.

What is the Objective of the study?
Specifically, what hypotheses are to be tested?
What is the variable of interest?
What unit is the treatment applied to?
What, exactly, is the unit measured?

Suppose an investigator wants to compare the oxygen levels in seven predefined "habitats" in the Louisiana marsh. He will randomly select and sample 4 sites in each habitat. One oxygen measurement is made at each site.

What **variable** is being measured? This variable will produce a series of measurements or quantities? **Oxygen levels** (usually in ppm)

What are the **treatments**? What is the investigator interested in comparing or testing for differences? **Habitats (t=7)**

Are there any **blocks** (i.e. sources of variation that should be recognized, but which are not important to the investigator). For example, did he replicate the experiment in several different rivers or several locations along the coast? Are the 4 "replicates" just multiple observations or are they taken in 4 separate places?
Apparently no blocks.

What are the **experimental units** for the experiment? What unit was the treatment applied to or what was sampled for each treatment (habitat)? **A site** (s=4)

Are there separate **sampling units** at each site, or is only one measurement taken in each experimental unit? In this case it is a water sample on which oxygen is measured. Since there is only a single sample at each site we can consider each sample to represent the site. Also the **sites**

If there were multiple samples taken at each site these would be the sampling units. These in turn can be split into sub-sampling units. Apparently this was not done.

Is that all? There are other issues, not all of which we have covered.

Are the treatments fixed or random?

Is the design balanced?

Are there any particular hypothesis tests of interest (contrasts)? Are the treatment levels quantitative?

Any other special post ANOVA applications? The topic of "Design" will be discussed in the second half of the course. For the moment our objective is only to learn to identify the components of an experiment.

Therefore, I will put a design description on the Internet and during each class period I expect you to have looked at it and to be prepared to answer the following questions.

Questions:

1) What is the treatment arrangement for this experiment?

(a) single factor (b) factorial (c) nested

2) What is the experimental design for this experiment?

(a) CRD (b) RBD (c) LSD (e) Split-plot (d) Repeated Measures

3) Does it seem more likely that the treatments are fixed or random?

(a) fixed (b) random

4) What is the experimental unit for this experiment?

(a) pens (b) diets (c) live weight (d) egg yolk weight (e) individual chickens

5) What is the sampling unit for this experiment?

(a) pens (b) diets (c) live weight (d) egg yolk weight (e) individual chickens

6) What is the dependent variable for this experiment?

(a) pens (b) diets (c) live weight (d) egg yolk weight (e) individual chickens

7) What is the treatment variable for this experiment?

(a) pens (b) diets (c) live weight (d) egg yolk weight (e) individual chickens

8) If the design is RBD, what are the blocks?

(a) pens (b) diets (c) live weight (d) egg yolk weight (e) individual chickens (f) NA

9 & 10) How many degrees of freedom are available for testing the treatment (combinations)?

Enter the correct value here: numerator = _____ , denominator = _____

For the little experiment discussed the source table is:

Source	d.f.
Treatment	**6**
Error	**21**
Total	27

A final note on the **Daily Designs**. These will start early in the semester with the simpler designs and progress to more complicated designs. Our only objective with these designs is that you learn to identify the important aspects and components of the designs. During the second half of the course we will discuss the designs in detail. We will see why these components exist, how they are analyzed and how they are interpreted.

At the start of each class I will roll a dice. If a value of 6 is obtained we will have a quiz. If any other value is rolled I will simply give you the answers to the quiz. If a value other than 6 is rolled, then that number is not to be counted again until after a 6 has occurred. For example, if I roll a 3 then any values of 3 on subsequent days would not count until a 6 occurred. Once a 6 is rolled all numbers are back in contention.

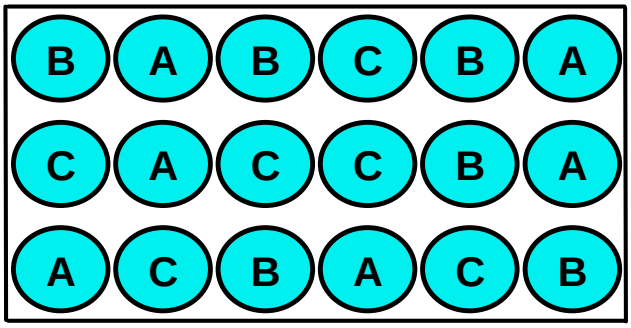
Identifying designs

The simplest type of design is the Completely Randomized Design (CRD)

This will consist of:

1) a **treatment** (at least one, possibly more), and
2) an **error term** (at least one, possibly more)

Example 1: A **researcher** is studying the **size of tomatoes** from plants grown under three watering **regimes**, (A) daily watering, (B) watering at 2–day intervals and (C) watering at 3–day intervals. Eighteen **plants** are planted in large pots and 6 are randomly selected for each watering regime. Plant productivity is measured as the **total weight** of the first 10 **tomatoes** (in cm) produced by each plant or pot (these are synonymous for this study since there is only one plant per pot).



Diagnosis: What are the treatments and experimental units?

The objective is to compare the total weight of tomatoes. The variable of interest (dependent variable) is the **total weight of the tomatoes**.

The **treatment** is the variable whose levels are to be compared. In this case the treatment has 3 levels, (i.e. the 3 watering regimes).

The **experimental unit** is the entity that is assigned a treatment. In this case the treatments are randomly assigned to pots / plants (synonymous in this case). Note that the weight values we analyze are NOT for individual tomatoes, but are the mean for each plant. Since we are interested in the mean for each plant we will have a dataset with only one value for each plant or pot, so we consider pot/plant to be the **sampling unit**. In this particular experiment the pots or plants are both the sampling unit and the experimental unit. In other experiments the tomatoes could be measured individually and the sampling unit would be individual tomatoes. However, when the individual units are summed or averaged to give just one measurement for each plant, the plant would be considered the sampling unit. Since the design does not have individual tomatoes we have the following.

Description: The model is $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

This analysis is a CRD with a single factor treatment and an experimental error. Each treatment was replicated 6 times.

The source table:

Source	d.f.	EMS
Watering Regime	t–1 = 2	$\sigma^2 + Q_\tau^2$
Plant (Regime)	t(n–1) = 15	σ^2
Total	tn–1 = 17	

Random versus Fixed treatment effects: There are a few other aspects of a design that will be important to identifying designs and achieving the correct analysis. One of these is to determine if treatment levels are randomly chosen from a large (theoretically infinite) number of choices, or if the treatment levels represent all possible levels, or at least all levels of interest.

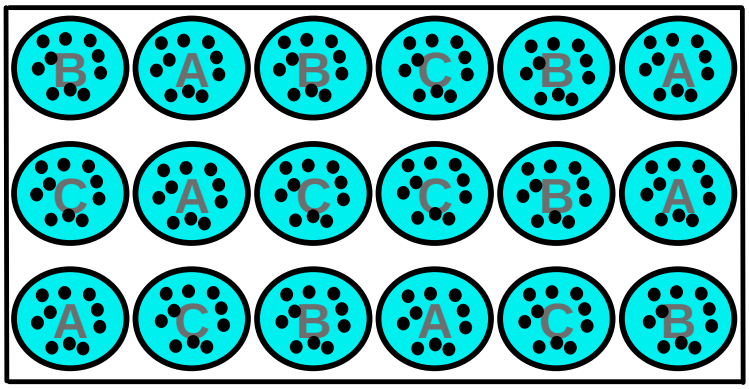
Treatment levels that are randomly chosen from a large number of possible levels estimate the variability among the levels, and are actually estimates of variance components. These are represented as σ_τ^2, and represent the variability among all the levels of the treatment. When the levels of the treatment represent all possible levels, or if they are chosen by the investigator as the only levels of interest they are called fixed effects, and statistical inference is limited to the treatment levels included in the experiment. Fixed effect treatments do not estimate variances, but rather the sum of squared treatment effects (deviations of each treatment level from the overall mean, i.e. $\tau_i^2 = \left(\bar{Y}_i - \bar{Y}_.\right)^2$).

These are summed ($\Sigma\tau_i^2 = \sum_{i=1}^{n}\left(\bar{Y}_i - \bar{Y}_.\right)^2$) and are represented as Q_τ^2

There are obviously many possible watering regimes, but the 3 of interest above do not appear to be "randomly chosen". They would be fixed.

For most of the simple analyses that we will examine the analytical differences between fixed effects and random effects will not be obvious. However, for larger experimental designs the differences become very important, so we will continue to pay attention to this aspect of each analysis.

Example 2: Suppose we modified the experiment above slightly by maintaining individual measurements for each tomato. Now, instead of 18 measurements (one mean per plant) we would have 180 measurements, one weight for each of 10 tomatoes on each of the 18 plants. How would this change the design?



Diagnosis: What are the treatments and experimental units?

The objective is to compare the weight of tomatoes, but in this case the variable of interest (dependent variable) is the **weight of individual tomatoes**.

The **treatment** is still the 3 watering regimes.

The **experimental unit** (the entity that is assigned a treatment) is still the pots or plants.

The **sampling unit** is now the individual tomatoes, since we now measure each tomato individually and record a separate measurement for each tomato fruit.

Description: The model is $Y_{ijk} = \mu + \tau_i + \gamma_{ij} + \varepsilon_{ijk}$

This analysis is a CRD with a single factor treatment and with both an experimental error for the experimental units and a sampling error for the sampling units. Each treatment was replicated 6 times and has 10 sampling units per experimental unit.

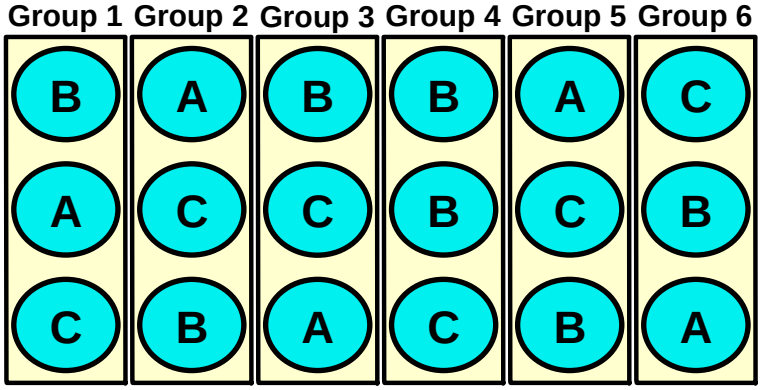
The source table:

Source	d.f.	EMS
Regime	t–1 = 2	$\sigma^2 + n\sigma^2_\gamma + Q^2_\tau$
Plant (Regime)	t(p–1) = 15	$\sigma^2 + n\sigma^2_\gamma$
Tomato (Plant x Regime)	tp(n–1) = 162	σ^2
Total	tpn–1 = 179	

Note that the σ^2 estimates the component of variability among tomatoes and σ^2_γ estimates the component of variability among the pots or plants. These two sources of variability (among tomatoes and among pots) may or may not differ (i.e. σ^2_γ may equal zero). Also note that the appropriate error term for the watering regime is the experimental error term.

Example 3: Suppose the room where we were to culture the plants has highly variable conditions. In particular there is a strong east to west variation in temperature and light conditions. If we place the treatments completely at random, we may get too many of one treatment on the east, where conditions are better, making that treatment look better than it really is. Also, the differing conditions will cause extra variability among the plants production and, therefore, the tomato weights. This extra variability will be added to the error term, reducing the power of the experiment (tests are the most powerful when the error term is small and the treatment differences are large). If we ignore the east-west variation it will become part of the error term.

So we decide to place our treatments in 6 groups. Each group contains only 1 replicate of the each treatment, but the 6 groups themselves constitute our replication. Note that the 6 groups could be any type of grouping that accounts for variation. For example, if we only had 3 suitable pots or chambers for the experiment we could replicate the experiment in 6 time periods. As before, in the first experiment, we will take the total weight of 10 tomatoes per plant for comparing the plants.



Diagnosis: What are the treatments and experimental units?

The objective is to compare tomato total weights. The variable of interest is the total weight of 10 tomatoes, as before. The treatment is still the 3 watering regimes. The experimental unit is still the pots or plants, and these also the sampling units which again provide one measurement for each plant.

The new wrinkle is the 6 different groups on which we conduct our experiment. Due to the potential differences (east-west), we will want to remove this variation from the error term. We do this by "blocking" on group and actually putting a separate variable in the model to account for block differences. If we don't, this additional variation among groups (east-west) goes into the error term and reduces our power.

Description: The model is $Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$

This analysis is a Randomized Block Design (RBD) with a single factor treatment and an experimental error.

The source table:

Source	d.f.	EMS
Block	b–1 = 5	$\sigma^2 + t\sigma^2_\beta$
Regime	t–1 = 2	$\sigma^2 + Q^2_\tau$
Block x Regime	(t–1)(b–1) = 10	σ^2
Total	tb–1 = 17	

Example 4: Suppose we now combine the second and third design. We conduct the experiment with only 3 pots in each of the 6 groups, but we measure and record 10 individual tomato weights from each plant.

Diagnosis: What are the treatments and experimental units?

The objective is to compare tomato weights, and we will again do this with individual tomatoes.

The treatment is still the 3 watering regimes.

The experimental unit is still the pots or plants, and the sampling units (individual tomatoes) again provide one measurement for each plant.

Again we are blocking on the groupings.

Description: The model is $Y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$

This analysis is a Randomized Block Design (RBD) with a single factor treatment and with both an experimental error and a sampling error.

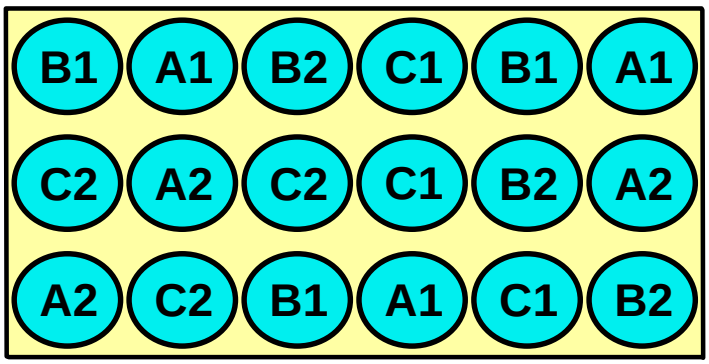
The source table:

Source	d.f.	EMS
Block	b–1 = 5	$\sigma^2 + n\sigma^2_{\beta\tau} + nt\sigma^2_\beta$
Regime	t–1 = 2	$\sigma^2 + n\sigma^2_{\beta\tau} + Q^2_\tau$
Block x Regime	(t–1)(b–1) = 10	$\sigma^2 + n\sigma^2_{\beta\tau}$
Tomato (Block x Regime)	tb(n–1) = 162	σ^2
Total	tbn–1 = 179	

Example 5: One last example. Suppose we go back to our first experiment, the CRD where we take the **total weight** of 10 tomatoes per plant. However, suppose we have two treatments of interest instead of the single factor (regime). The two treatments of interest are (a) the 3 watering regimes and we are also interested in comparing between plants that receive a low level of fertilizer and those that receive a high level. We will examine all 6 combinations of treatments, and 3 of the 18 potted plants will be allocated to each treatment combination.

Notice that the treatments are cross classified. Each level of "Watering" occurs with each level of "Fertilizer" such that all possible combinations exist. This is characteristic of a factorial analysis of variance, also called a two-way ANOVA.

	Watering regime treatment (3 levels)		
Fertilizer treatment (2 levels)	daily watering (A)	2-day intervals (B)	3-day intervals (C)
Low (1)	3 replicate pots	3 replicate pots	3 replicate pots
High (2)	3 replicate pots	3 replicate pots	3 replicate pots



Diagnosis: What are the treatments and experimental units?

The objective is to compare tomato weights, and we will have one mean per plant.

There are now two treatments, one with 2 levels and one with 3 levels. This is called a 2x3 factorial treatment arrangement. Note that "low fertilizer" and "high fertilizer" would appear to cover all possible levels of this treatment, so this treatment is also fixed.

Both the experimental units and sampling units are still the pots / plants.

There is no blocking in this experiment, so this analysis is a CRD with a 2x3 factorial treatment arrangement.

Description: The model is $Y_{ijk} = \mu + \tau_{1i} + \tau_{2j} + \tau_1\tau_{2ij} + \varepsilon_{ijk}$

The source table:

Source	d.f.	EMS
Fertilizer	$t_1-1 = 1$	$\sigma^2 + Q^2_{\tau 1}$
Regime	$t_2-1 = 2$	$\sigma^2 + Q^2_{\tau 2}$
Fertilizer x Regime	$(t_1-1)(t_2-1) = 2$	$\sigma^2 + Q^2_{\tau 1\tau 2}$
Pot (Fertilizer x Regime)	$t_1t_2(n-1) = 12$	σ^2
Total	$t_1t_2n-1 = 17$	

Alternatively: let "treatments" represent all t=6 combinations of t_1 by t_2

Source	d.f.
Treatments	t–1 = 5
Experimental Error	t(n–1) = 12
Total	tn–1 = 17

Some final notes on the identification of designed experiments.

There are other types of designs that we will discuss this semester. We will discuss the Latin Square Design (LSD), which has a peculiar structure with two sources of blocking (generically referred to as "rows" and "columns"). The treatments are arranged in such a way that each treatment occurs once in each row and once in each column.

Another major class of designs will be discussing under the title of "Split plots and Repeated measures". This class of designs has an initial structure that can be CRD, RBD or LSD with treatment arrangements that may be either a single factor or factorial. Then each experimental unit is either split into two or more units and a new treatment applied to the sub-units of the experimental unit, or each experimental unit is sampled over time. In the latter case "time" becomes a source of variation of interest, and is included in the source table.

For example, in our tomato plant example above we might split our experimental unit (the plant) into two levels, the lower half of the plant and the upper half of the plant. We might take 10 tomatoes from the lower half and 10 tomatoes from the upper half. The new variable "level" would be included in the experiment. This would be an example of a split plot. Another possibility is that we are interested in the changing size of the tomatoes produced by each plant over time. We might take the first 10 tomatoes (time 1), and the second 10 tomatoes (time 2), etc. This would give a new variable called time that would show if the size of tomatoes was the same over time or not, and if the different watering regimes were similar of different over time.

There is also another treatment arrangement called the nested treatment arrangement. This arrangement has a hierarchical arrangement of treatments with one level nested within the higher level. Do not confuse this treatment arrangement with the nested error terms discussed above where the sampling error is also nested within the experimental error in a hierarchical fashion. Also note that it is possible to have many levels of nesting of both treatments and error terms.

Random effects versus fixed effects. When the effects are fixed we are usually interested in the individual treatment levels. Frequently, we will calculate a mean for each treatment level and do comparisons and tests among those means. Random effects, however, represent a random selection from a very large number of possible levels, and we are not usually interested in each of the individual levels selected. For random effects we are most likely interested in the overall mean and the variability about that mean, so we would likely want to place a confidence interval on that mean.

Introduction

Major topics (a comprehensive outline is provided elsewhere)

- Regression : SLR, Multiple, Curvilinear & Logistic
- Experimental Design : CRD, RBD, LSD, Split-plot & Repeated Measures
- Treatment arrangements : Single factor, Factorial, Nested

Course Objectives

The objectives of the introductory course were to develop an understanding of elemental statistics, the ability to understand and apply basic statistical procedures. We will develop those concepts further, applying the terminology and notation from the basic methods courses to advanced techniques for making statistical inferences.

We will cover the major methodologies of parametric statistics used for prediction and hypotheses testing (primarily regression and experimental design).

Our emphasis will be on RECOGNIZING analytical problems and on being able to do the statistical analysis with SAS software. We will see SAS programs and output for virtually all analyses covered this semester.

Daily Design

I will be placing a design description on the Internet for each class. You should plan on examining this design before class. At the beginning of each class I will randomly determine whether we have a quiz on that design or not. I do not intend to spend much time on this daily activity. If there is a quiz, I will allow 5 minutes for you to answer and turn in quiz. If not, I will give you the answers.

A quiz will consist of specifying the dependent variable, experimental and sampling units, treatments, blocks, random effects, etc.

We will address most of these in the design section of the course (following regression). However, some will be covered in the daily design.

Notes on Exams

I usually schedule a review session late on Tuesday for the Thursday exams and Sunday for Tuesday exams. Review session is entirely voluntary, and you may leave anytime. I will have not plan on covering material. I plan only to answer questions. There will be no review for the final exam.

On the exam you will be allowed to bring a **calculator**

- I do not expect to have many calculations on the exam, but there may be some.
- For example, calculating a t-test for a slope for an hypothesized value other than zero (thought this may be in the output, always check first).
- Also confidence intervals on slopes and treatment means.

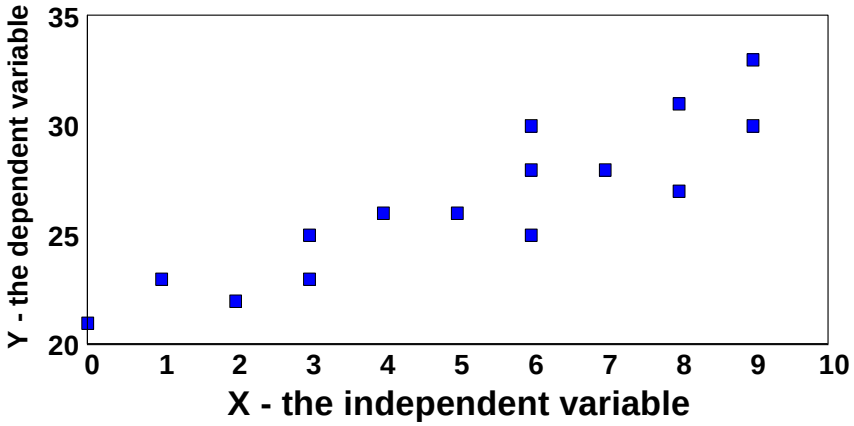
You may also bring an 8.5 by 11 inch sheet of paper with equations or whatever else you wish to include. You may write on both sides of that piece of paper.

I will provide you with these on an exam. You will need to understand MY t-tables. See interment for copies of these tables.

All exams, including the final, will be in our regular classroom(s).

Simple Linear Regression (review?)

The objective: Given points plotted on two coordinates, Y and X, find the best line to fit the data.



The concept: Data consists of paired observations with a presumed potential for the existence of some underlying relationship. We wish to determine the nature of the relationship and quantify it if it exists.

> Note that we cannot prove that the relationship exists by using regression (i.e. we cannot prove cause and effect). Regression can only show if a "correlation" exists, and provide an equation for the relationship.

Given a data set consisting of paired, quantitative variables, and recognizing that there is variation in the data set, we will define,

POPULATION MODEL (SLR): $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

> This is the model we will fit. It is the equation describing straight line for a population and we want to estimate the parameters in the equation. The population parameters to be estimated are for the underlying model, $\mu_{y.x} = \beta_0 + \beta_1 X_i$, are:
>
> - $\mu_{y.x}$ = the true population mean of Y at each value of X
> - β_0 = the true value of the Y intercept
> - β_1 = the true value of the slope, the change in Y per unit of X

Terminology

Dependent variable: variable to be predicted

> Y = dependent variable (all variation occurs in Y)

Independent variable: predictor or regressor variable

> X = independent variable (X is measured without error)

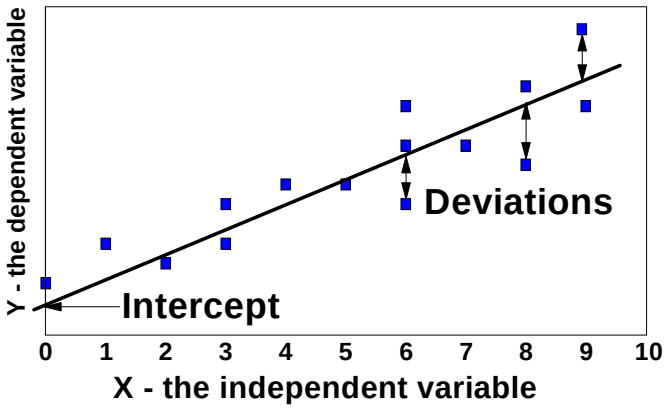
Intercept: value of Y when X = 0, point where the regression line passes through the Y axis. The units on the intercept are the same as the "Y" units

Slope: the value of the change in Y for each unit increase in X. The units on the slope are "*Y*" units per "*X*" unit

Deviation: distance from an observed point to the regression line, also called a residual.

Least squares regression line: the line that minimizes the squared distances from the line to the individual observations.

Regression line



The regression line itself represents the mean of Y at each value of X ($\mu_{y.x}$).

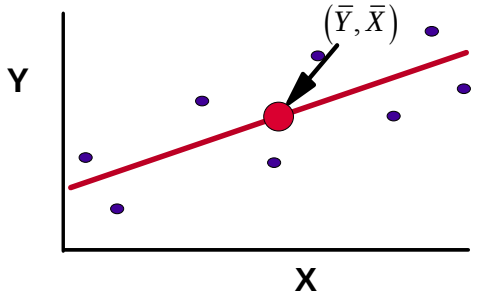
Regression calculations

All calculations for simple linear regression start with the same values. These are,

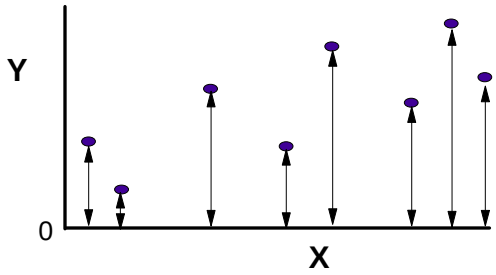
$$\sum_{i=1}^{n} X_i,\ \sum_{i=1}^{n} X_i^2,\ \sum_{i=1}^{n} Y_i,\ \sum_{i=1}^{n} Y_i^2,\ \sum_{i=1}^{n} X_i Y_i,\ n$$

Calculations for simple linear regression are first adjusted for the mean. These are called "corrected values". They are corrected for the MEAN by subtracting a "correction factor".

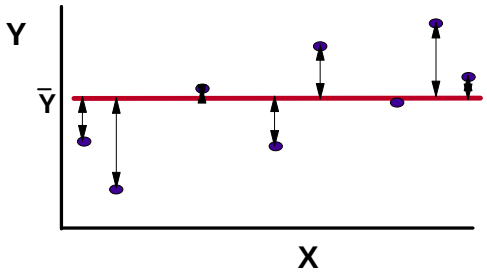
As a result, all simple linear regressions are adjusted for the mean of *X* and *Y* and pass through the point $\left(\overline{Y}, \overline{X}\right)$.



The original sums and sums of squares of *Y* are distances and squared distances from zero. These are referred to as "uncorrected" meaning unadjusted for the mean.



The "corrected" deviations sum to zero (half negative and half positive) and the sums of the squares are squared distances from the mean of Y.



Once the means $\left(\bar{X},\bar{Y}\right)$ and corrected sums of squares and cross products $\left(S_{XX},S_{YY},S_{XY}\right)$ are obtained, the calculations for the parameter estimates are:

$$\text{Slope} = b_1 = S_{XY} \Big/ S_{XX}$$

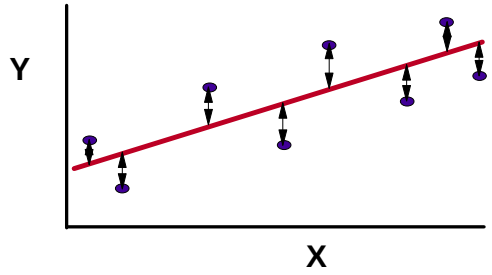
$$\text{Intercept} = b_0 = \bar{Y} - b_1\bar{X}$$

We have fitted the sample equation

$Y_i = b_0 + b_1X_i + e_i$, which estimates the population parameters of the model, $Y_i = \beta_0 + \beta_1X_i + \varepsilon_i$

Variance estimates for regression

After the regression line is fitted, variance calculations are based on the deviations from the regression. From the regression model $Y_i = b_0 + b_1X_i + e_i$ we derive the formula for the deviations $e_i = Y_i - \left(b_0 + b_1X_i\right)$ or $e_i = Y_i - \hat{Y}_i$.



As with other calculations of variance, we calculate a sum of squares (corrected for the mean). This is simplified by the fact that the deviations, or residuals, already have a mean of zero,

$$\text{SSResiduals} = \sum_{i=1}^{n} e_i^2 = \text{SSError}.$$

The degrees of freedom (d.f.) for the variance calculation is n–2, since two parameters are estimated prior to the variance (β_0 and β_1).

The variance estimate is called the MSE (Mean square error). It is the SSError divided by the d.f.,

$$MSE = SSE \Big/ \left(n-2\right).$$

The variances for the two parameter estimates and the predicted values are all different, but all are based on the MSE, and all have n–2 d.f. (t-tests) or n–2 d.f. for the denominator (F tests).

$$\text{Variance of the slope} = MSE \Big/ S_{XX}$$

Variance of the intercept = $MSE\left(\frac{1}{n}+\frac{-\bar{X}^2}{S_{XX}}\right)$

Variance of a predicted value at X_i = $MSE\left(\frac{1}{n}+\frac{\left(X_i-\bar{X}\right)^2}{S_{XX}}\right)$

Any of these variances can be used for a t-test of an estimate against an hypothesized value for the appropriate parameter (i.e. slope, intercept or predicted value respectively).

ANOVA table for regression

A common representation of regression results is an ANOVA table. Given the SSError (sum of squared deviations from the regression), and the initial total sum of squares (S_{YY}), the sum of squares of Y adjusted for the mean, we can construct an ANOVA table

Simple Linear Regression ANOVA table

	d.f.	Sum of Squares	Mean Square	F
Regression	1	SSRegression	MSReg	$MSReg/MSError$
Error	n–2	SSError	MSError	
Total	n–1	S_{YY} = SSTotal		

In the ANOVA table

The SSRegression and SSError sum to the SSTotal, so given the total (S_{YY}) and one of the two terms, we can get the other.

The easiest to calculate first is usually the SSRegression since we usually already have the necessary intermediate values.

SSRegression = $(S_{XY})^2 / S_{XX}$

The SSRegression is a measure of the "improvement" in the fit due to the regression line. The deviations start at S_{YY} and are reduced to SSError. The difference is the improvement, and is equal to the SSRegression.

This gives another statistic called the R^2. What portion of the SSTotal ($\underline{S_{YY}}$) is accounted for by the regression?

R^2 = SSRegression / SSTotal

The degrees of freedom in the ANOVA table are,

n–1 for the total, one lost for the correction for the mean (which also fits the intercept)

n–2 for the error, since two parameters are estimated to get the regression line.

1 d.f. for the regression, which is the d.f. for the slope.

The F test is constructed by calculating the MSRegression / MSError. This F test has 1 in the numerator and (n–2) d.f. in the denominator. This is exactly the same test as the t-test of the slope against zero.

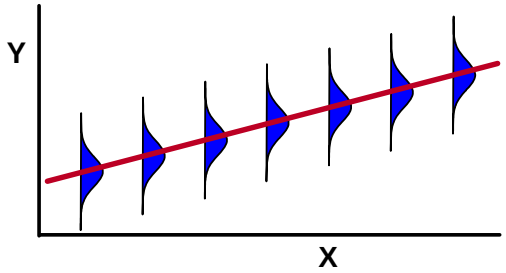
To test the slope against an hypothesized value (say zero) using the t-test with n–2 d.f., calculate

$$t = \frac{b_1 - b_{1Hypothesized}}{S_{b_1}} = \frac{b_1 - 0}{\sqrt{MSE / S_{XX}}}$$

Assumptions for the Regression

We will recognize 4 assumptions

1) Normality – We take the deviations from regression and pool them all together into one estimate of variance. Some of the tests we use require the assumption of normality, so these deviations should be normally distributed.



For each value of X there is a population of values for the variable Y (normally distributed).

2) Homogeneity of variance – When we pool these deviations (variances) we also assume that the variances are the same at each value of X_i. In some cases this is not true, particularly when the variance increases as X increases.

3) X is measured without error! Since variances are measured only vertically, all variance is in Y, no provisions are made for variance in X.

4) Independence. This enters in several places. First, the observations should be independent of each other (i.e. the value of e_i should be independent of e_j, for $i \neq j$). Also, in the equation for the line $Y_i = b_0 + b_1 X_i + e_i$ we assume that the term e_i is independent of the rest of the model. We will talk more of this when we get to multiple regression.

So the four assumptions are:

- Normality
- Homogeneity of variance
- Independence
- X measured without error

These are explicit assumptions, and we will examine or test these assumptions when possible. There are also some other assumptions that I consider implicit. We will not state these, but in some cases they can be tested. For example,

- There is order in the Universe. Otherwise, what are you investigating? Chaos?
- The underlying fundamental relationship that I just fitted a straight line to really is a straight line. Sometimes this one can be examined statistically.

Characteristics of a Regression Line

- The line will pass through the point $\left(\overline{Y}, \overline{X}\right)$ (also the point b_0, 0)
- The sum of deviations will be zero ($\Sigma e_i = 0$)
- The sum of squared deviations (measured vertically, $\Sigma e_i^2 = \Sigma\left(Y_i - b_0 - b_1 X_i\right)^2$ of the points from the regression line will be a minimum.
- Values on the line can be described by the equation $\hat{Y}_i = b_0 + b_1 X_i$.
- The line has some desirable properties (if the assumptions are met)
  - $\mathrm{E}\left(b_0\right) = \beta_0$
  - $\mathrm{E}\left(b_1\right) = \beta_1$
  - $\mathrm{E}\left(\overline{Y}_X\right) = \mu_{Y.X}$

  Therefore, the parameter estimates and predicted values are unbiased estimates.
- Note that linear regression is considered statistically robust. That is, the tests of hypothesis tend to give good results if the assumptions are not violated to a great extent.

Crossproducts and correlation

Crossproducts are used in a number of related calculations (can be + or –).

- a crossproduct = $Y_i X_i$
- Sum of crossproducts = $\Sigma Y_i X_i$
- Corrected sum of crossproducts = S_{XY}
- Covariance = $S_{XY} \Big/ \left(n-1\right)$
- Slope = $S_{XY} \Big/ S_{XX}$
- SSRegression = $S_{XY}^2 \Big/ S_{XX}$
- Correlation = $S_{XY} \Big/ \sqrt{S_{YY} S_{XX}}$
- $R^2 = r^2 = S_{XY}^2 \Big/ S_{YY} S_{XX}$ = SSRegression / SSTotal

Simple Linear Regression Summary

- See Simple linear regression notes from EXST7005 for additional information, including the derivation of the equations for the slope and intercept. You are not responsible for these derivations.
- Know the terminology, characteristics and properties of a regression line, the assumptions, and the components to the ANOVA table.
- You will not be fitting regressions by hand, but I will expect you to understand where the values on SAS output come from and what they mean.
- Particular emphasis will be placed on working with, and interpreting, numerical regression analyses. Analyses will mostly be done with SAS.