A factorial is a way of entering two or more treatments into an analysis.

The description of a factorial usually includes a measure of size, a 2 by 2, 3 by 4, 6 by 3 by 4, 2 by 2 by 2, etc.

Interactions were discussed.

Interactions test additivity of the main effects

Interactions are a measure of inconsistency in the behavior or the cells relative to the main effects.

Interactions are tested along with the main effects

Interactions should not be ignored if significant.

Factorial analyses can be done as two-way ANOVAs in SAS, or they can be done as contrasts.

The Randomized Block Design

This analysis is similar in many ways to a "two-way" ANOVA

The CRD is defined by the linear model, $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$. The simplest version of the CRD has one treatment and one error term. The factorial treatment arrangement discussed previously occurred within a CRD, and it had several different treatments, $Y_{ijk} = \mu + \tau_{1i} + \tau_{2j} + \tau_{1i}\tau_{2j} + \varepsilon_{ijk}$.

This model has two treatments and one error. It could have many more treatments, and it would still be a factorial design. Designs having a single treatment or multiple treatments can all occur within a CRD and are referred to as different treatment arrangements.

There are other modifications of a CRD that could be done. Instead of multiple treatments we may find it necessary to subdivide the error term.

Why would we do this? Perhaps there is some variation that is not of interest. If we ignore it, that variation will go to the error term. For example, suppose we had a large agricultural experiment, and had to do our experiment in 8 different fields, or due to space limitations in a greenhouse experiment we had to separate our experiment into 3 different greenhouses or 5 different incubators. Now there is a source of variation that is due to different fields, or different greenhouses or incubators!

If we do it as a CRD, we put our treatments in the model, but if there is some variation due to field, greenhouse or incubator it will go to the error term. This would inflate our error term an make it more difficult to detect a difference (we would lose power).

How do we prevent this? First, make sure each treatment occurs in each field, greenhouse or incubator (preferably balanced). Then we would factor the new variation out of the error term by putting it in the model.

$$Y_{ijk} = \mu + \beta_i + \tau_j + \beta_i\tau_j + \varepsilon_{ijk}$$

This is not a new treatment. We will call it a BLOCK. This looks like a factorial, but it is not because the blocks are not a source of variation that we are interested in discussing.

Also, in a factorial the interaction term is likely to be something of interest. In a block design the interaction is an error term, representing random variation of experimental units across treatments.

Another difference, treatments can be either fixed or random. If both treatments are fixed, the interaction is fixed. However, blocks are usually random, and the block interaction is always random.

So why are we blocking?

It is usually used to add replication to an experiment. Additional replicates are added in another field, another greenhouse. On the one hand, the larger experiment should add power. On the other hand, if we do not take measures to keep the new variation out of the error term, we may lose power due to the larger error.

So, how does this affect our analysis?

We still have treatments with the test of treatments in the ANOVA (an F test).

We can still do post-hoc tests on the treatments.

There is only one new issue, the error term. To examine this we will need to look at the expected mean squares (EMS) for the Randomized Block Design.

RBD EMS

We will examine two possible types of models.

In the first model we have treatments and blocks and nothing else. Each treatment occurs in each block ONCE. The experiment is similar to a factorial in some regards, but not many.

The model is

$$Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}$$

In this model the error term (ε_{ij}) actually comes from the block by treatment interactions ($\beta\tau_{ij}$). This is the only error available, but that is OK. It is usually a good error term because it represents random variation among the experimental units.

Blocks \ Treatments	**A1**	**A2**	**A3**
Block 1	$\mathbf{a_1b_1}$	$\mathbf{a_2b_1}$	$\mathbf{a_3b_1}$
Block 2	$\mathbf{a_1b_2}$	$\mathbf{a_2b_2}$	$\mathbf{a_3b_2}$
Block 3	$\mathbf{a_1b_3}$	$\mathbf{a_2b_3}$	$\mathbf{a_3b_3}$

This looks like a factorial.

The analysis is the same as the factorial, we get marginal sums or means and proceed to calculate the SS for blocks and treatments and "interaction" as before.

However, there is one big difference. If this was a factorial we would have Treatment A, Treatment B and the A*B interaction.

What would you use as an error term? We would not have one. A factorial ANOVA must have an error term for testing treatments and interaction.

However, since the "interaction" in a block design is assumed to be random variation among experimental units, it serves as an error term.

So the model works for Block designs.

$$Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}$$

The "interaction" term is a useful and respectable error term.

We do however, in this case, have one additional assumption.

Assume that there is "no interaction between the treatments and blocks". By interaction here we mean that the treatment patterns are the same in each block. We do not have a treatment behaving one way in one block, and behaving differently in another block.

So the term represents random variation in experimental units and not some interaction in the same sense as "interaction" in a factorial design.

So what about those EMS?

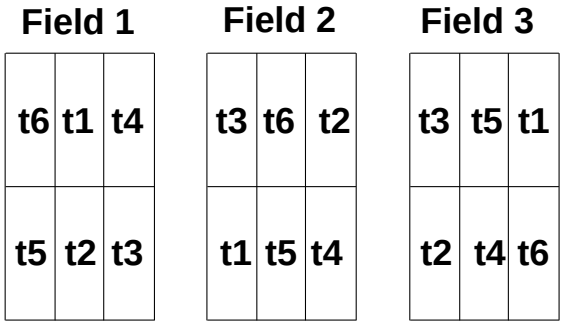
For the CRD we had two cases

Source	d.f.	EMS Random	EMS Fixed
Treatment	t–1	$\sigma^2_\varepsilon + n\sigma^2_\tau$	$\sigma^2_\varepsilon + n\,\Sigma\tau_i^2 / t-1$
Error	t(n–1)	σ^2_ε	σ^2_ε
Total	tn–1		

For the Block design we have two cases, one with just blocks and treatments, and one with replicate observations within the cells.

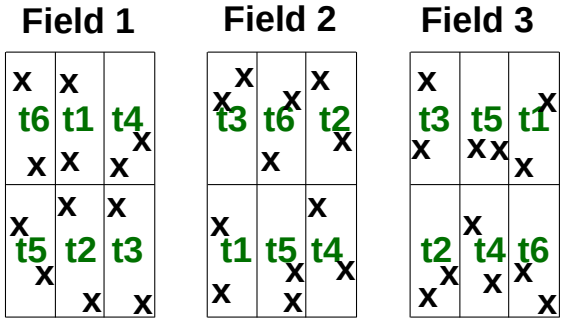
Source	d.f.	EMS (no reps)	EMS (with Reps)
Treatment	t–1	$\sigma^2_\varepsilon + b\sigma^2_\tau$	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + nb\sigma^2_\tau$
Block	b–1	$\sigma^2_\varepsilon + t\sigma^2_\beta$	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + nt\sigma^2_\beta$
Exptl Error	(b–1)(t–1)	σ^2_ε	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta}$
Rep Error	tb(n–1)		σ^2_ε
Total	tbn–1		

What is the nature of the replicates within the block by treatment cells?



Suppose the experimental unit is a plot in a field.
- We are evaluating plant height.
- The treatment to be compared is 6 varieties of soy beans.
- The experiment is done in 3 fields (blocks).
- The error term is the field by variety combinations.
- This experiment is unreplicated within blocks.

Additional replication is usually done in one of two ways.
- If we have only one plot (experimental unit) for each treatment in each field, we could sample several times within each plot, sampling plant height at several places in the plot.
- Our "sampling unit" is a smaller unit than the experimental unit (a plot) so we have sampling error.



Replicated within blocks as multiple samples in an experimental unit. Error is sampling error.

Another type of error comes from having several plots with a given soybean variety in each plot.

Here each variety of soybean has several experimental units in each field.

In this case the additional replication represents a second experimental error, one for block by treatment combinations and one for replicate plots within a block.

In this case we have replicated experimental units in each block.

Field 1

t5	t3	t2
t1	t2	t6
t6	t5	t1
t4	t4	t3

Field 2

t3	t5	t1
t4	t3	t2
t6	t5	t1
t4	t2	t6

Field 3

t1	t4	t2
t6	t5	t4
t5	t1	t3
t3	t6	t2

Factorial EMS

I haven't mentioned factorials EMS.

Developing EMS can be pretty simple. Start with the lowest unit, and move up the source table adding additional variance components for each new term.

Source	EMS with Reps
Treatment	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + nb\sigma^2_\tau$
Block	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + nt\sigma^2_\beta$
Exptl Error	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta}$
Rep Error	σ^2_ε

Interaction components occur on their own line, and on the source line for each higher effect contained in the interaction.

Each main effect gets its own source.

Now consider whether the effects are fixed or random. Modify fixed effects to show SSEffects instead of variance components.

Source	EMS with Reps
Treatment	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + nb\,{\Sigma\tau_i^2}/{t-1}$
Block	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta} + nt\sigma^2_\beta$
Exptl Error	$\sigma^2_\varepsilon + n\sigma^2_{\tau\beta}$
Rep Error	σ^2_ε

If the model is an RBD we're done, because the interaction is always a random variable.

For factorials that are random models and mixed models were done.

Consider what the F test should be for the treatment. Surprise, SAS always uses the residual error term!

But for factorials there is one last detail. It is perfectly possible in factorial designs that both effects are fixed, and if both effects are fixed the interaction is also fixed!

Source	EMS with Reps
Treatment A	$\sigma_\varepsilon^2 + nb\,\Sigma\tau_{Ai}^2 \big/ (a-1)$
Treatment B	$\sigma_\varepsilon^2 + na\,\Sigma\tau_{Bi}^2 \big/ (b-1)$
Interaction A*B	$\sigma_\varepsilon^2 + n\sum(\tau_A\tau_B)_{ij}^2 \big/ (a-1)(b-1)$
Error	σ_ε^2

And a FIXED effect occurs only on its own line, no other!! The fixed interaction disappears from the main effects!!!

Now what is the error term for testing treatments and interaction? Maybe SAS is right? Or maybe SAS just doesn't know what is fixed and what is random.

Testing ANOVAs in SAS

So tell SAS what is random and what is fixed.

Look for the following additions to SAS program.

How do we tell SAS which terms to test with what error term?

How do we get SAS to output EMS?

How do we get SAS to automagically test the right treatment terms with the right error terms?

Summary

Randomized Block Designs modify the model by factoring a source of variation out of the error term in order to reduce the error variance and increase power. If the basis for blocking is good, this will be effective. If the basis for blocking is not good, we lose a few degrees of freedom from the error term and may actually lose power.

The block by treatment combinations (interaction?) provide a measure of variation in the experimental units and provide an adequate error term.

We have an additional assumption that this error term represents ONLY experimental error, and not some real interaction between the treatments and blocks.

Expected mean squares for the RBD indicate that the experimental error term is the correct error term, whether there is a sampling unit or not.

Factorial designs, where effects are random or mixed are similar to RBD EMS. THE TREATMENT INTERACTION IS ACTUALLY USED AS AN ERROR TERM!

When the treatments are fixed, the main effects do not contain the interaction term, and the residual error term is the appropriate error term.

Sample size in ANOVA

Some textbooks use a slightly different expression for the equation, but it is the same as the equation discussed previously. One minor change is the expression $n \geq (t_{\alpha/2} + t_{\beta})^2 \frac{S^2}{\overline{d}^2}$. An alternative to the use of $\overline{d}$ is the expressing of the difference as a percentage of the mean. For example, if we wanted to test for a difference that was 10% of the mean we could use the expression $n \geq (t_{\alpha/2} + t_{\beta})^2 \frac{S^2}{\left(0.1\overline{Y}\right)^2}$. This expression can further be altered to express the difference in terms of the coefficient of variation $CV = S/\overline{Y}$. Calculating the sample sized needed to detect a 10% change in the mean then becomes $n \geq (t_{\alpha/2} + t_{\beta})^2 (10CV)^2$.

In analysis of variance we may also want to be able to detect a certain difference between two means (μ_1 and μ_2) out of the treatment means we are studying, so our difference will be $\mu_1 - \mu_2$. A prior analysis, or a pilot study, may provide us with an estimate of the variance (MSE in ANOVA). From here we can use a formula pretty much the same as for the t-test discussed earlier. There is one other little detail, however.

We are basically testing H_0: $\mu_1^2 - \mu_2^2 = \delta$, from the 2 sample t-test. Recall from our linear combinations we have a variance for this linear combination that is the sum of the individual variances of the mean. Therefore, the variance will be. $\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$. Since we are usually pooling variances (ANOVA) then the formula simplifies to $MSE(\frac{1}{n_1} + \frac{1}{n_2})$

. Furthermore, since we usually attempt to have balanced experiments (equal sample size in each group) for analysis of variance the formula further simplifies to an expression similar to one seen previously, except for the addition of "2", $2MSE/n$. The additional "2" occurs when we are testing for difference in two means (H_0: $\mu_1 = \mu_2$) as opposed to testing a mean against an hypothesized value (H_0: $\mu = \mu_0$).

Note one very important thing here. In this formula "n" represents each group or population being studied, that is, each "treatment level" in an analysis of variance! So for ANOVA or two-sample t-test with equal variance and equal n, the expression for sample size is $n \geq \frac{2(t_{\alpha/2} + t_{\beta})^2 \text{MSE}}{\overline{d}^2}$. Note that this "n" is for each treatment. In a two sample t-test, each population would have a sample size of "n", so the total number of observations would be 2n.

In ANOVA we have "t" treatments; each would have a sample size of "n", so the total number of observations would be tn. How often are we likely to have situations with equal variance and equal n? Is this realistic? Actually, yes it is.

First, ANOVA traditionally required equal variances, though more modern analytical techniques can address the lack of homogeneity. If necessary, equal variances may be achieved by a transformation or some other fix. If variance is nonhomogeneous you could use the larger estimates and get a conservative estimate of "n".

Second, the most common application for sample size calculation is in planning NEW studies, and of course in planning new studies you usually do not PLAN on unbalanced designs and non homogeneous variance.

So these situations are realistic.

Summary

Finally we saw that this formula is applicable to two-sample t-tests and ANOVA, with some modifications in the estimate of the variance. These modifications are the same ones needed for the 2-sample t-test as dictated by our study of linear combinations. However, the calculations are simplified by the common ANOVA assumption of equal variance and the prevalence of balanced experiments.

Review of Analysis of Variance procedures.

1) H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots = \mu_t = \mu$

2) H_1: some μ_i is different

3a) Assume that the observations are normally distributed about each mean, or that the residuals (i.e. deviations) are normally distributed.

b) Assume that the observations are independent

c) Assume that the variances are homogeneous

4) Set the level of type I error. Usually $\alpha = 0.05$

5) Determine the critical value. For a balanced CRD with a single factor treatment the test is an F test with t–1 and t(n–1) degrees of freedom ($F_{\alpha=0.05,\ t-1,\ t(n-1)\ d.f.}$).

6) Obtain data and evaluate.

The treatment sum of squares, as developed by Fisher, are converted to a "variance" and tested with an F test against the pooled error variance. In practice, the sum of squares are usually calculated and presented with the degrees of freedom in a table called an ANOVA table. For a balanced design (all n_i equal) the calculations are,

The uncorrected SS for treatments is $USS_{Treatments} = \dfrac{\sum_{i=1}^{t}\left(\sum_{j=1}^{n} Y_{ij}\right)^2}{n} = n\sum_{i=1}^{t}\left(\dfrac{\sum_{j=1}^{n} Y_{ij}}{n}\right)^2$.

The uncorrected SS for the total is $SS_{Total} = \sum_i \sum_j Y_{ij}^2$

The correction factor for both terms is $CF = \dfrac{\left(\sum_i^t \sum_j^n Y_{ij}\right)^2}{tn}$

Our ANOVA analyses will be done with PROC MIXED and PROC GLM. There is a PROC ANOVA, but it is a subset of PROC GLM.

LSMeans calculation

The calculations of LSMeans are different. For a balanced design, the results will be the same. However, for unbalanced designs the results will often differ.

The MEANS statement in SAS calculates a simple mean of all available observations in the treatment cells.

The LSMeans statement will calculate the mean of the treatment cell means.

Example:

The MEAN of 4 treatments, where the observations are 3,4,8 for a1, 3,5,6,7,9 for a2, 7,8,6,7 for a3 and 3,5,7 for a4 is 5.8667.

The individual cells means are 5, 6, 7 and 5 for a1, a2, a3 and a4 respectively. The mean of these 4 values is 5.75. This would be the LSMean.

Raw means

Treatments	a1	a2	a3	Means
b1	5	6	9	6.5
	7	8		
		4		
b2	7	5	5	6.6
	9		7	
Means	7	5.75	7	

LSMeans means

Treatments	a1	a2	a3	Means
b1	6	6	9	7
b2	8	5	6	6.33
Means	7	5.5	7.5	

Confidence Intervals on Treatments

Like all confidence intervals on normally distributed estimates, this will employ a t-value and will be of the form Mean $\pm\ t_{\alpha/2} S_{\bar{Y}}$

The treatment mean can be obtained from a means (or LSMeans) statement, but the standard deviation provided is not the correct standard error for the interval.

The standard error in a simple CRD with fixed effects is the square root of *MSE/n*, where *n* is the number of observations used in calculating the mean.

The calculation requires other considerations when random components are involved. For example, in PROC MIXED the use of the Satterthwaite and Kenward-Roger approximations, the use of various estimation methods (the default is REML) and specifications of covariance structure are all things that can affect degrees of freedom.

The use of *MSE* in the numerator is the default in PROC GLM, and if a different error is desired it must be specified by the user. PROC MIXED is capable of detecting and using and error other than the *MSE* where appropriate.

If there are several error terms (e.g. experimental error and sampling error) use the one that is appropriate for testing the treatments. When an error term other than the residual is appropriate for testing the treatments, the degrees of freedom for the tabular t value are the d.f. from the error term used for testing. This variance term would also be used to calculate the standard error for treatment means.

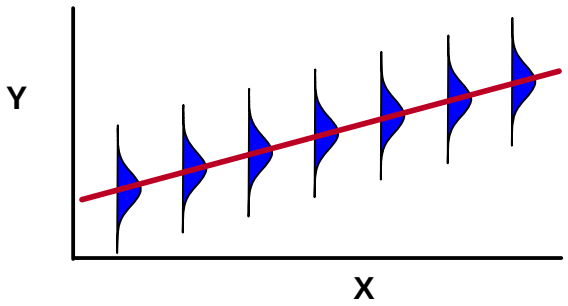
Simple Linear Regression

Simple regression applications are used to fit a model describing a linear relationship between two variables. The aspects of least squares regression and correlation were developed by Sir Francis Galton in the late 1800's.

The application can be used to test for a statistically significant correlation between the variables. Finding a relationship does not prove a "cause and effect" relationship, but the model can be used to quantify a relationship where one is known to exist. The model provides a measure of the rate of change of one variable relative to another variable..

There is a potential change in the value of variable *Y* as the value of variable *X* changes.

Variable values will always be paired, one termed an independent variable (often referred to as the *X* variable) and a dependent variable (termed a *Y* variable). For each value of *X* there is assumed to be a normally distributed population of values for the variable *Y*.



The linear model which describes the relationship between two variables is given as

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

The "Y" variable is called the dependent variable or response variable (vertical axis).

$\mu_{y.x} = \beta_0 + \beta_1 X_i$ is the population equation for a straight line. No error is needed in this equation because it describes the line itself. The term $\mu_{y.x}$ is estimated with at each value of X_i with $\hat{Y}$.

$\mu_{y.x}$ = the true population mean of *Y* at each value of *X*

The "X" variable is called the independent variable or predictor variable (horizontal axis).

β_0 = the true value of the intercept (the value of *Y* when *X* = 0)

β_1 = the true value of the slope, the amount of change in *Y* for each unit change in *X* (i.e. if *X* changes by 1 unit, *Y* changes by β_1 units).

The two population parameters to abe estimated, β_0 and β_1 are also referred to as the regression coefficients.

All variability in the model is assumed to be due to Y_i, so variance is measured vertically

The variability is assumed to be normally distributed at each value of X_i

The X_i variable is assumed to have no variance since all variability is in Y_i (this is a new assumption)

The values β_0 and β_1 (b_0 and b_1 for a sample) are called the regressions coefficients.

The β_0 value is the value of Y at the point where the line crosses the Y axis. This value is called the intercept. If this value is zero the line crosses at the origin of the X and Y

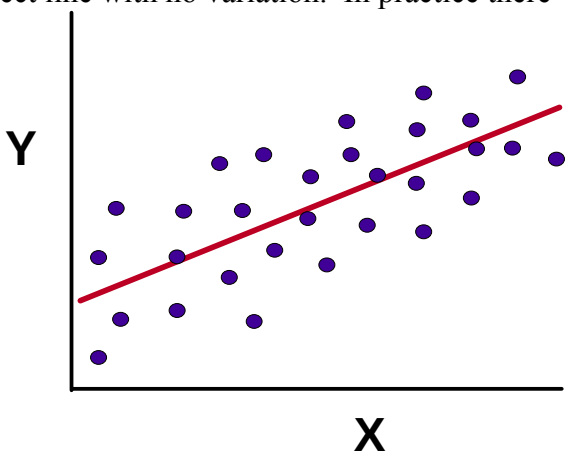
axes, and the linear equation reduces from "$Y_i=b_0+ b_1X_i$" to "$Y_i=b_1X_i$" and is said to have "no intercept", even though the regression line does cross the Y axis. The units on b_0 are the same units as for Y_i.

The β_1 value is called the slope. It determines the incline or angle of the regression line. If the slope is 0, the line is horizontal. At this point the linear model reduced to "$Y_i=b_0$", and the regression is said to have "no slope". The slope gives the change in Y per unit of X. The units on the slope are then "Y units per X unit".

The population equation for the line describes a perfect line with no variation. In practice there is always variation about the line. We include an additional term to represent this variation.

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ for a population

$Y_i = b_0 + b_1 X_i + e_i$ for a sample



When we put this term in the model, we are describing individual points as their position on the line plus or minus some deviation

The Sum of Squares of deviations from the line will form the basis of a variance for the regression line

When we leave the e_i off the sample model we are describing a point on the regression line, predicted from the sample estimates. To indicate this we put a "hat" on the Y_i value, $\hat{Y}_i = b_0 + b_1 X_i$.

Characteristics of a Regression Line

The line will pass through the point $\bar{Y}, \bar{X}$ (also the point b_0, 0)

The sum of squared deviations (measured vertically) of the points from the regression line will be a minimum.

Values on the line for any value of X_i can be described by the equation $\hat{Y}_i = b_0 + b_1 X_i$

Common objectives in Regression : there are a number of possible objectives

Determine if there is a relationship between Y_i and X_i .

This would be determined by some hypothesis test.

The strength of the relationship is, to some extent, reflected in the correlation or R^2 value.

Determine the value of the rate of change of Y_i relative to X_i .

This is measured by the slope of the regression line.

This objective would usually be accompanied by a test of the slope against 0 (or some other value) and/or a confidence interval on the slope.

Establish and employ a predictive equation for Y_i from X_i .

This objective would usually be preceded by a Objective 1 above to show that a relationship exists.

The predicted values would usually be given with their confidence interval, or the regression with its confidence band.

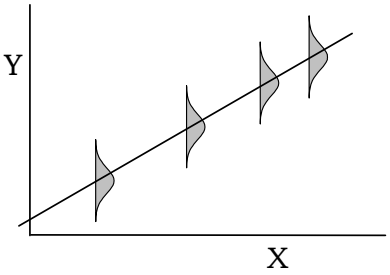
Assumptions in Regression Analysis

Independence

The best guarantee of this assumption is random sampling. This is a difficult assumption to check.

This assumption is made for all tests we will see in this course.

Normality of the observations at each value of X_i (or the pooled deviations from the regression line)



- This is relatively easy to test if the appropriate values are tested (e.g. residuals in ANOVA or Regression, not the raw Y_i values). This can be tested with the Shapiro-Wilks W statistic in PROC UNIVARIATE.
- This assumption is made for all tests we have seen this semester except the Chi square tests of Goodness of Fit and Independence

Homogeneity of error (homogeneous variances or homoscedasticity)

- This is easy to check for and to test in analysis of variance (S^2 on mean or tests like Bartalett's in ANOVA). In Regression the simplest way to check is by examining the the residual plot.
- This assumption is made for ANOVA (for pooled variance) and Regression. Recall that in 2 sample t-tests the equality of the variances need not be assumed, it can be readily tested.

X_i measured without error: This must be assumed in ordinary least squares regressions, since all error is measured in a vertical direction and occurs in Y_i .

Assumptions – general assumptions

The *Y* variable is normally distributed at each value of *X*

The variance is homogeneous (across *X*).

Observations are independent of each other and e_i independent of the rest of the model.

Special assumption for regression.

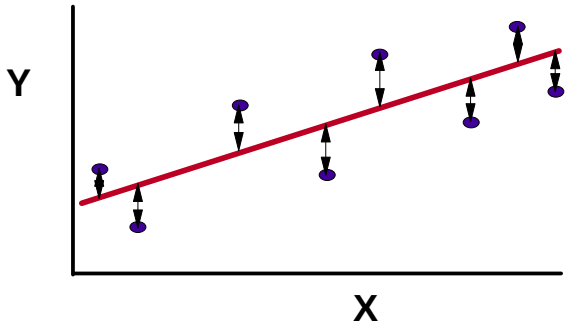
Assume that all of the variation is attributable to the dependent variable (*Y*), and that the variable *X* is measured without error.

Note that the deviations are measured vertically, not horizontally or perpendicular to the line.

Fitting the line

Fitting the line starts with a corrected SSDeviation, this is the SSDeviation of the observations from a horizontal line through the mean.

The line will pass through the point $\bar{X}$, $\bar{Y}$. The fitted line is pivoted on this point until it has a minimum SSDeviations.



How do we know the SSDeviations are a minimum? Actually, we solve the equation for e_i, and use calculus to determine the solution that has a minimum of the sum of squared deviations.

$$Y_i = b_0 + b_1 X_i + e_i$$

$$e_i = Y_i - (b_0 + b_1 X_i) = Y_i - \hat{Y}_i$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

The line has some desirable properties

$$E(b_0) = \beta_0$$

$$E(b_1) = \beta_1$$

$$E(\bar{Y}_X) = \mu_{Y.X}$$

Therefore, the parameter estimates and predicted values are unbiased estimates.

Derivation of the formulas

You do not need to learn this derivation for this class! However you should be aware of the process and its objectives.

Any observation from a sample can be written as $Y_i = b_0 + b_1 X_i + e_i$.

where; e_i = a deviation of the observed point from the regression line

The idea of regression is to minimize the deviation of the observations from the regression line, this is called a Least Squares Fit. The simple sum of the deviations is zero, $\Sigma e_i = 0$, so minimizing will require a square or an absolute value to remove the sign.

The sum of the squared deviations is,

$$\sum e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum \left(Y_i - b_0 - b_1 X_i\right)^2$$

The objective is to select b_0 and b_1 such that Σe_i^2 is a minimum, by using some techniques from calculus. We have previously defined the uncorrected sum of squares and corrected sum of squares of a variable Y_i.

The corrected sum of squares of *Y*

The uncorrected SS is $\sum Y_i^2$

The correction factor is $\left(\sum Y_i\right)^2 \Big/ n$

The corrected SS is $CSS = S_{YY} = \sum\left(Y_i - \bar{Y}\right)^2 = \sum Y_i^2 - \left(\sum Y_i\right)^2 \Big/ n$

We will call this corrected sum of squares S_{YY} and the correction factor C_{YY}

The corrected sum of squares of *X*

We could define the exact same series of calculations for X_i, and call it S_{XX}

The corrected cross products of *Y* and *X*

We need a cross product for regression, and a corrected cross product. The cross product is X_iY_i.

The uncorrected sum of cross products is $\sum Y_i X_i$

The correction factor for the cross products is $C_{XY} = \left(\Sigma Y_i\right)\left(\Sigma X_i\right) \Big/ n$

The corrected cross product is $CCP = S_{XY} = \sum\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right) = \sum Y_i X_i - \dfrac{\left(\Sigma Y_i\right)\left(\Sigma X_i\right)}{n}$

The formulas for calculating the slope and intercept can be derived as follows

Take the partial derivative with respect to each of the parameter estimates, b_0 and b_1.

For b_0 :

$\dfrac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial b_0} = 2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)(-1)$, which is set equal to 0 and solved for b_0.

$-\Sigma Y_i + nb_0 + b_1\Sigma X_i = 0$ (this is the first "normal equation")

Likewise, for b_1 we obtain the partial derivative, set it equal to 0 and solved for b_1.

$\dfrac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial b_1} = 2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)(-X_i)$

$-\Sigma(Y_iX_i - b_0X_i - b_1X_i^2) = -\Sigma Y_iX_i + b_0\Sigma X_i + b_1\Sigma X_i^2)$ (second "normal equation")

The normal equations can be written as,

$$b_0 n + b_1\Sigma X_i = \Sigma Y_i$$
$$b_0\Sigma X_i + b_1\Sigma X_i^2 = \Sigma Y_i X_i$$

At this point we have two equations and two unknowns so we can solve for the unknown regression coefficient values b_0 and b_1.

For b_0 the solution is: $nb_0 = \Sigma Y_i - b_1 \Sigma X_i$ and $b_0 = \Sigma Y_i / n - b_1 \Sigma X_i / n = \bar{Y}_i - b_1 \bar{X}_i$.

Note that estimating β_0 requires a prior estimate of b_1 and the means of the variables X and Y.

For b_1, given that, $b_0 = \Sigma Y_i / n - b_1 \Sigma X_i / n$ and $\Sigma Y_i X_i = b_0 \Sigma X_i + b_1 \Sigma X_i^2$ then

$$\Sigma Y_i X_i = \left(\Sigma Y_i / n - b_1 \Sigma X_i / n \right) \Sigma X_i + b_1 \Sigma X_i^2 = \Sigma Y_i \Sigma X_i / n - b_1 \left(\Sigma X_i \right)^2 / n + b_1 \Sigma X_i^2$$

$$\Sigma Y_i X_i - \Sigma Y_i \Sigma X_i / n = b_1 X_i^2 - b_1 \left(\Sigma X_i \right)^2 / n = b_1 \left(\Sigma X_i^2 - \left(\Sigma X_i \right)^2 / n \right)$$

$$b_1 = \frac{\Sigma Y_i X_i - \Sigma Y_i \Sigma X_i / n}{\Sigma X_i^2 - \left(\Sigma X_i \right)^2 / n} = \frac{S_{YX}}{S_{XX}}$$ so b_1 is the corrected cross products over the corrected sum of squares of X

The intermediate statistics needed to solve all elements of a SLR are ΣY_i, ΣX_i, ΣY_i^2, ΣX_i^2, $\Sigma Y_i X_i$ and n. We have not seen ΣY_i^2 used in the calculations yet, but we will need it later to calculate variance.

Review

We want to fit the best possible line through some observed data points. We define this as the line that minimizes the vertically measured distances from the observed values to the fitted line.

The line that achieves this is defined by the equations

$$b_0 = \Sigma Y_i / n - b_1 \Sigma X_i / n = \bar{Y}_i - b_1 \bar{X}_i$$

$$b_1 = \frac{\Sigma Y_i X_i - \Sigma Y_i \Sigma X_i / n}{\Sigma X_i^2 - \left(\Sigma X_i \right)^2 / n} = \frac{S_{YX}}{S_{XX}}$$

These calculations provide us with two parameter estimates that we can then use to get the equation for the fitted line. $\hat{Y}_i = b_0 + b_1 X_i$.

Testing hypotheses about regressions

The total variation about a regression is exactly the same calculation as the total for Analysis of Variance. SSTotal = SSDeviations from the mean = Uncorrected SSTotal – Correction factor

The simple regression analysis will produce two sources of variation.

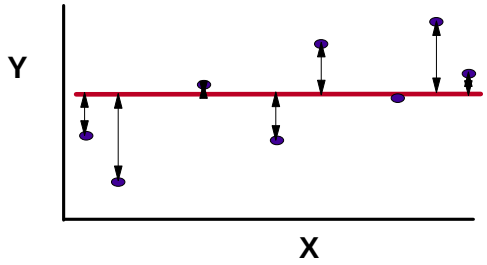
SSRegression – the variation explained by the regression

SSError – the remaining, unexplained variation about the regression line.

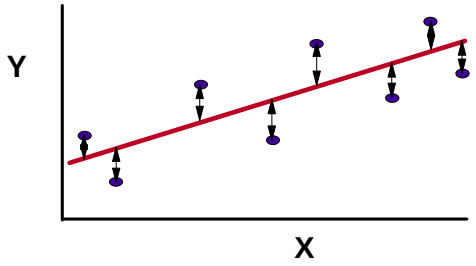
These sources of variation are expressed in an ANOVA source table.

Source	d.f.	
Regression	1	d.f. used to fit slope
Error	n–2	error d.f.
Total	n–1	d.f. lost in adjusting for ("correcting for") the mean

Note that one degree of freedom is lost from the total for the "correction for the mean", which actually fits the intercept. The single regression d.f. is for fitting the slope.



The correction fits a flat line through the mean



The "regression" actually fits the slope.

The difference between these two models is that one has no slope, or a slope equal to zero ($b_1 = 0$) and the other has a slope fitted. Testing for a difference between these two cases is the common hypothesis test of interest in regression and it is expressed as $H_0: \beta_1 = 0$.

The results of a regression are expressed in an ANOVA table. The regression is tested with an F test, formed by dividing the MSRegression by the MSError.

Source	df	SS	MS	F
Regression	1	SSRegression	MSRegression	$MSRegression/MSError$
Error	n – 2	SSError	MSError	
Total	n – 1	SSTotal		

This is a one tailed F test, as it was with ANOVA, and it has 1 and n–1 d.f. It tests the null hypothesis $H_0: \beta_1 = 0$ versus the alternative $H_1: \beta_1 \neq 0$.

The R^2 statistic

This is a popular statistic for interpretation. The concept is that we want to know what proportion of the corrected total sum of squares is explained by the regression line.

Source	d.f.	SS
Regression	1	SSReg
Error	n–2	SSError
Total	n–1	SSTotal

In the regression the process of fitting the regression the SSTotal is divided into two parts, the sum of squares "explained" by the regression (SSRegression) and the remaining

unexplained variation (SSError). Since these sum to the SStotal, we can calculate what fraction of the total was fitted or explained the regression. This is often expressed as a percentage of the total sum of squares explained by the model, and is given by R^2 = SSRegression / SSTotal.

This is often multiplied by 100% and expressed as a percent.

We might state that the regression explains 75% of the total variation.

This is a very popular statistic, but it can be very misleading.

For some studies an R^2 value of 25% or 35% can be pretty good. For example, if you are trying to relate the abundance of an organism to environmental variables. On the other hand, if you are doing mophometric relationships, like relating a crabs width to its length, an R^2 value of less than 90% is pretty bad.

A note on regression models applied to transformed variables.

Studies of mophometric relationships, including relationships of lengths to weights, should be done with logarithmic values of both X and Y. The log(Y) on log(X) model, called a power model, is a very flexible model used for many purposes.

Many other models involving logs, powers, inverses are possible. These will fit curves of one shape or another. When using transformed variables in regression, all tests and confidence intervals are placed on the transformed values. Otherwise, they are used like any other simple linear regression.

Numerical Example : Some freshwater-fish ectoparasites accumulate on the fish as it grows. Once the parasite is on the fish, it does not leave. The parasite completes it's live cycle after the fish is consumed by a bird and finds it way again into the water. Since the parasite attaches and does not leave, *older fish should accumulate more parasites*. We want to test this hypothesis.

Raw data with squares and crossproducts

Observation	Age	Parasites	Age^2	Parasite^2	Age*Parasite
1	1	3	1	9	3
2	2	7	4	49	14
3	3	8	9	64	24
4	3	12	9	144	36
5	3	10	9	100	30
6	4	15	16	225	60
7	4	14	16	196	56
8	5	16	25	256	80
9	6	17	36	289	102
10	6	15	36	225	90
11	6	16	36	256	96
12	7	19	49	361	133
13	7	21	49	441	147
14	8	18	64	324	144
15	9	17	81	289	153
16	9	20	81	400	180

Summary data

Sum	83	228	521	3628	1348
Mean	5.1875	14.25	32.5625	226.75	84.25
n	16	16	16	16	16

Intermediate Calculations

$\Sigma X = 83$ $\Sigma Y = 228$

$\Sigma X^2 = 521$ $\Sigma Y^2 = 3628$

Mean of X_i = $\bar{X} = 5.1875$ Mean of Y_i = $\bar{Y} = 14.25$

$\Sigma XY = 1348$ $n = 16$

Correction factors and Corrected values (Sums of squares and crossproducts)

CF for X	$C_{xx} = 430.5625$	Corrected SS X	$Sxx = 90.4375$
CF for Y	$C_{yy} = 3249$	Corrected SS Y	$Syy = 379$
CF for XY	$Cxy = 1182.75$	Corrected CP XY	$S_{xy} = 165.25$

ANOVA Table (values needed): SSTotal = 379

SSRegression = $165.25^2 / 90.4375 = 301.9495508$

SSError = 379 – 301.9495508 = 77.05044921

Source	df	SS	MS	F
Regression	1	301.9495508	301.9495508	54.8639723
Error	14	77.05044921	5.503603515	
Total	15	379.		Tabular $F_{0.05;\ 1,\ 14} = 4.600$
				Tabular $F_{0.01;\ 1,\ 14} = 8.862$

Model Parameter Estimates

$$\text{Slope} = b_1 = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}_.\right)\left(X_i - \bar{X}_.\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}_.\right)^2} = \frac{S_{xy}}{S_{xx}} = 165.25 / 90.4375 = 1.827228749$$

Intercept = b_0 = $\bar{Y} - b_1 \bar{X} = 14.25 - 1.827228749 * 5.1875 = 4.771250864$

Regression Equation $Y_i = b_0 + b_1 * X_i + e_i = Y_i = 4.771250864 + 1.827228749 * X_i + e_i$

Regression Line $\hat{Y}_i = b_0 + b_1 * X_i = Y_i = 4.771250864 + 1.827228749 * X_i$

$$\text{Standard error of } b_1: S_{b_1} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}\left(X_i - \bar{X}_.\right)^2}} = \sqrt{\frac{MSE}{S_{xx}}} \quad \text{so } S_{b_1} = \sqrt{\frac{5.5036}{90.4375}} = 0.2467$$

Confidence interval on b_1 where $b_1 = 1.827228749$ and $t_{(0.05/2,\ 14df)} = 2.145$

$P(1.827228749 - 2.145*0.246688722 \le \beta_1 \le 1.827228749 + 2.145*0.246688722) = 0.95$

$P(1.29808144 \le \beta_1 \le 2.356376058) = 0.95$

Testing b_1 against a specified value: e.g. $H_0: \beta_1 = 5$ versus $H_1: \beta_1 \neq 5$

where $b_1 = 1.827228749$, $S_{b1} = 0.246688722$ and $t_{(0.05/2,\ 14df)} = 2.145$

$= (1.827228749 - 5) / 0.246688722 = -12.86144$

Standard error of the regression line (i.e. $\hat{Y}_i$) : $s_{\mu_{\hat{Y}|X}} = \sqrt{MSE\left(\frac{1}{n}+\frac{(X_i-\overline{X}_.)^2}{\sum_{i=1}^{n}(X_i-\overline{X}_.)^2}\right)}$

Standard error of the individual points (i.e. Y_i): This is a linear combination of $\hat{Y}_i$ and e_i, so the variances are the sum of the variance of these two, where the variance of e_i is MSE. The standard error is then $s_{\mu_{Y|X}} = \sqrt{s^2_{\mu_{\hat{Y}|X}} + MSE} =$

$$\sqrt{MSE\left(\frac{1}{n}+\frac{(X_i-\overline{X}_.)^2}{\sum_{i=1}^{n}(X_i-\overline{X}_.)^2}\right)+MSE} = \sqrt{MSE\left(1+\frac{1}{n}+\frac{(X_i-\overline{X}_.)^2}{\sum_{i=1}^{n}(X_i-\overline{X}_.)^2}\right)}$$

Standard error of b_0 is the same as the standard error of the regression line where $X_i = 0$

Square Root of [5.503603515 (0.0625 + 26.91015625/90.4375)] = 1.407693696

Confidence interval on b_0, where $b_0 = 4.771250864$ and $t_{(0.05/2,\ 14df)} = 2.145$

$P(4.771250864 - 2.145*1.407693696 \le \beta_0 \le 4.771250864+2.145*1.407693696) = 0.95$
$P(1.751747886 \le \beta_0 \le 7.790753842) = 0.95$

Estimate the standard error of an individual observation for number of parasites for a ten-year-old fish: $\hat{Y} = b_0 + b_1 X_i = 4.77125 + 1.82723*X = 4.77125 + 1.82723*10 = 23.04354$

Square Root of [$5.503603515*(1+0.0625+(10 - 5.1875)^2/90.4375)$] =
Square Root of [$5.503603515*(1+0.0625+(23.16015625)/90.4375)$] = 2.693881509

Confidence interval on $\mu_{Y|X=10}$

$P(23.04353836 - 2.145*2.693881509 \le \mu_{Y|X=10} \le 23.04353836+2.145*2.693881509) = 0.95$
$P(17.26516252 \le \mu_{Y|X=10} \le 28.82191419) = 0.95$

Calculate the coefficient of Determination and correlation

$R^2 =$ 0.796700662 or 79.67006617 %
$r =$ 0.892580899

See SAS output

Overview of results and findings from the SAS program