

PubH 7405: REGRESSION ANALYSIS



ONE-FACTOR EXPERIMENT DESIGNS

Designed experiments are conducted to **“demonstrate”** a **cause-and-effect relation** between one or more explanatory factors (or predictors) and a response variable. The demonstration of a cause-and-effect relationship is accomplished, to put it in a simple way, by altering the level or levels of the explanatory factors (i.e. “designed”) and observing the effect of the changes (i.e. designed values of predictors X’s) on the response variable Y. Designed experiments are often used as **“comparative”** in natures.

A Simple Example:

An experiment on the effect of Vitamin C on the prevention of colds could be simply conducted as follows. A number of n children (the sample size) are randomized; half were each give a 1,000-mg tablet of Vitamin C daily during the test period and form the “experimental group”. The remaining half , who made up the “control group” received “placebo” – an identical tablet containing no Vitamin C – also on a daily basis. At the end, the “Number of colds per child” could be chosen as the outcome/response variable, and the means of the two groups are compared.

Pay attention to the Explanatory variable (Predictor), Factor levels or treatment arms, Experimental units, and Outcome/Response variable.

Assignment of the treatments (factor levels: Vitamin C or Placebo) to the experimental units (children) was performed using a process called “randomization”. The purpose of randomization was to “balance” the characteristics of the children in each of the treatment groups, so that the difference in the response variable, the number of cold episodes per child, can be rightly attributed to the effect of the predictor – the difference between Vitamin C and Placebo.

The simplest form of designed experiments is the “completely randomized design” where treatments are randomly assigned to the experimental units – regardless of their characteristics. This design is most useful when the experimental units are relatively homogeneous with respect to known confounders. Otherwise, heterogeneous experimental units are divided into homogeneous “block”; and randomizations of treatments are carried out within each block. The result would be a “randomized complete block design”; the analyses are different, say, **One-way ANOVA versus Two-way ANOVA.**

INFERENCES & VALIDITIES

- Two major levels of inferences are involved in interpreting a study
 - ❖ The first level concerns Internal validity; the degree to which the investigator draws the correct conclusions about what actually happened in the study.
 - ❖ The second level concerns External Validity (also referred to as generalizability or inference); the degree to which these conclusions could be appropriately applied to people and events outside the study.

External Validity

Internal Validity

Truth in
The Universe

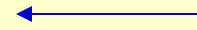
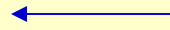
Truth in
The Study

Findings in
The Study

Research Question

Study Plan

Study Data



With the goal of maximizing the validity of the inferences, the investigator reverses the process: **(i) designs a study plan in which the choice of the research question, the subjects, and the measurements enhances the External Validity,** **(ii) is conducive to implementation with a high degree on Internal Validity.**

That is to focus on the External Validity first (Design) then Internal Validity (Implementation).

Statistical contributions involve both Internal Validity (for example, helping to select a sensitive “endpoint”) and External Validity (helping to choose a proper design and an adequate sample size).

THE BASIC ISSUE

Most of the times, inexperienced researchers mistakenly act like there is an identifiable, existent parent population or populations of subjects. We act as if the sample or samples is/are obtained from the parent population or populations according to a carefully defined technical procedure called random sampling. And we simply compare population means.

This is not true in real-life biomedical studies. The laboratory investigator uses animals in his projects but the animals are not randomly selected from any large population of animals. The clinician, who is attempting to describe the results he has obtained with a particular therapy, cannot say that his patients is a random sample from a parent population of patients.

THE VALUE OF TRIALS

- Because they are not population-based (there is not an identifiable, existent parent population of subjects for sample selection), biomedical studies – designed experiments are “**comparative**”. That is the validity of the conclusions is based on a **comparison**.
- In a clinical trial, we compare the results from the “treatment group” versus the results from the “placebo group”. The validity of the comparison is backed by the randomization.

As an introduction, we cover only the basic form: Completely Randomized Design. There are many other forms; some of the more important are:

Randomized Complete Block Design;

Factorial Design;

Repeated-measure Design;

Cross-over Design

Each is conducted differently and data from each are analyzed differently

DATA ANALYSIS METHODS

Two-sample t-test: to compare two population means;

One-way ANOVA (Analysis Of Variance) to compare several population means;

ANCOVA (Analysis Of Covariance): to compare two or several population means **adjusted** for one or more confounders

Analysis A: COMPARISON OF TWO POPULATION MEANS

- In this type of problems, we have two independent samples (n_1, \bar{y}_1, s_1^2) and (n_2, \bar{y}_2, s_2^2) ; the n 's being the sample sizes, \bar{y} the sample means, and s^2 the sample variances (the s are standard deviations).
- Often called the “**two-sample problem**”
- Considered as samples with population means μ_1 and μ_2
- **The aim is to compare the two population means.**
- (“Y” is the “response”, a measure of interest)

#1: TWO-SAMPLE t-TEST

- The Null Hypothesis considered is $H_0: \mu_1 = \mu_2$ or equivalently, $H_0: \mu_2 - \mu_1 = 0$.
- The assumptions are:
 - ❖ Independent observations
 - ❖ Two Normal Distributions
 - ❖ Variances are equal
- (Normal assumption may be dropped if sample sizes are large – due to Central Limit Theorem)

#2: REGRESSION APPROACH

- Pool the data, treat Y as dependent variable
- Binary independent variable ($X=0/1$; for group1/2)
- The assumptions (on Y ; Regression of Y on X):
 - ❖ Independent observations
 - ❖ Normal Distribution for Y
 - ❖ Constant Variance
- **Same as assumptions of the two-sample t-test**

Normal Error Regression Model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

Independent Variable X is binary :

$$\mathbf{E(Y | X = x) = \beta_0 + \beta_1 x}$$

$$E(Y | x = 0; \text{group 1}) = \beta_0$$

$$E(Y | x = 1; \text{group 2}) = \beta_0 + \beta_1$$

$$E(Y | x = 1) - E(Y | x = 0) = \beta_1$$

Same Null Hypothesis :

$$(\mu_2 = \mu_1) \Leftrightarrow (\mu_2 - \mu_1 = 0) \Leftrightarrow (\beta_1 = 0)$$

EQUIVALENCY

- **Same Assumptions**
- **Same Null Hypothesis**
- **In order to prove that they are the same t-test, at $df = (n-2)$; $n = n_1 + n_2$ we prove that they have the same test statistic too.**

$$t = \frac{\bar{y}_2 - \bar{y}_1}{\sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{\bar{y}_2 - \bar{y}_1}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Two-sample t-test

Regression's
test for independence

$$b_1 = \bar{y}_2 - \bar{y}_1$$

$$SE(b_1) = \sqrt{\frac{MSE}{\sum (x - \bar{x})^2}}$$

$$= s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$t = \frac{b_1}{SE(b_1)}$$

Conclusion:

We can do regression with a binary independent variable; the results are equivalent to those of the two-sample t-test to compare two population means.

AN ALTERNATIVE CODING

The **common practice** is to use “binary coding” (0/1); an alternative (0/1) coding is using (-1/+1).

With the alternative (-1/+1) coding the numerical value of the regression coefficient get cut in halves but all statistical decisions remain unchanged (because, for example, the standard error is also reduced accordingly).

Analysis B: COMPARISON OF POPULATION SEVERAL MEANS

- **Suppose we want to know whether there are differences in the means of more than two independent groups. For example, do families of different ethnic groups have different income levels?**
- **Two Questions here: How to measure the “difference”? and how to decide if the observed difference is real ? (i.e. statistically significant).**

QUESTION #1:

How do we measure the “difference” among several means? By subtraction? Which from which? If not, then what else can we “measure” differences? That is what should we use instead of difference $(x_2 - x_1)$ (or their ratio)?

QUESTION #2:

How do we decide if the “difference” among several sample means is large enough to conclude that the population means are different? That is what do we use instead of the t-test?

MULTIPLE COMPARISONS

- When you compare all possible pairs of means, there are more work; but that's not the problem
- To perform many tests increases the probability that one or more of the comparisons will result in a Type I error (true null hypothesis is wrongly rejected); e.g., suppose the null hypothesis is true and we perform 100 tests---each has a 0.05 probability of resulting in a Type I error; then 5 of these 100 tests would be statistically significant as the results of Type I errors. **Every time we do more than one, then the probability of Type I errors exceeds 0.05**

#1: ONE-WAY “ANOVA”

- What is needed is a different way to summarize the differences between several means and a method of simultaneously comparing these means in one step. This method is ANOVA or One-way ANOVA, for “ANalysis Of VAriance”.
- If that “one-step” test, the ANOVA F-test, is significant indicating that some pair(s) of means are different then we can start looking for that/those pairs- with “allowance” for multiple comparisons.

COMPONENTS OF TOTAL VARIATION

- The total variation in the combined sample can be decomposed into two components as follows:

$$(x_{ij} - \bar{x}) = (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x}):$$

- (1) The first term reflects the **variation within the samples**; the following sum is called the “**within sum of squares**”:
- $$SSW = \sum_{i,j} (x_{ij} - \bar{x}_i)^2 = \sum_i (n_i - 1) s_i^2$$

- (2) The difference, **SSB = SST - SSW**, is called the “**between sum of squares**” which measures the differences between samples:

$$SST = \sum_{i,j} (x_{ij} - \bar{x})^2$$

$$SSB = \sum_{i,j} (\bar{x}_i - \bar{x})^2 = \sum_i n_i (\bar{x}_i - \bar{x})^2$$

ANOVA: ANALYSIS OF VARIANCE

- **SST** measures the “total variation” in the combined sample with $(n-1)$ degrees of freedom, $n=\sum n_i$ is the total size. It is decomposed into:
 $SST=SSW+SSB$
- **SSW** measures the variation within samples with $\sum(n_i-1)=(n-k)$ degrees of freedom, and
- **SSB** measures the variation between sample means with $(k-1)$ degrees of freedom; $k=\#$ of groups

ANSWER #1:

SSB measures the variation, or difference, between sample means:

$$SSB = \sum_{i,j} (\bar{x}_i - \bar{x})^2 = \sum_i n_i (\bar{x}_i - \bar{x})^2$$

(which is a concept similar to the “variance”: variation among sample means)

“ANOVA” TABLE

- The breakdowns of the total sum of squares and its associated degree of freedom are displayed in the form of an “analysis of variance table” (ANOVA table) as follows:

Source of Variation	SS	df	MS	F ratio	p-val
Between samples	SSB	k-1	MSB	MSB/MSW	
Within samples	SSW	n-k	MSW		
Total	SST	n-1			

- MSW** is a natural extension of the pooled estimate s_p^2 as used in the two-sample t-test; It is a measure of the average variation within the k samples.

APPROACH TO QUESTION #2:

COMPARE the “average gap/difference” between sample means (MSB) to the average gap/difference between measurements in samples (MSW):

Use $F = MSB/MSW$

THE “F” TEST

- The test statistic F for the Analysis of Variance compares MSB (the average variation between the k sample means) and MSE (the average variation within the k samples), a value near 1 supports the null hypothesis of no differences between the k population means.
- If we apply the new method of “One-way ANOVA” to compare the means of two groups, the result is identical to that of a two-sided two-sample t -test

ANOVA ASSUMPTIONS

- **The Null Hypothesis considered is**
 $H_0: \mu_1 = \mu_2 = \dots = \mu_k$
- **The assumptions are:**
 - ❖ **Independent observations**
 - ❖ **k Normal Distributions**
 - ❖ **Variances are equal**
- **(Normal assumption may be dropped if sample sizes are large – due to Central Limit Theorem)**

#2: REGRESSION APPROACH

- Pool data, label measurement as “Y” (instead of “X”) and treat Y as dependent variable
- Binary independent variables ($X_i = 0/1$; for group k/i – choosing “group k” as “baseline” – it doesn’t matter which group is chosen)
- The assumptions (on Y; Regression of Y on X’s):
 - ❖ Independent observations
 - ❖ Normal Distribution for Y
 - ❖ Constant Variance
- Same as assumptions of one-way ANOVA

Typical Case : 3 groups, 2 indicators X 's

Normal Error Regression Model :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$\varepsilon \in N(0, \sigma^2)$$

Example: SBP Versus RACES

$$Y = SBP$$

$$X_1 = \begin{cases} 1 & \text{if White} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if Black} \\ 0 & \text{otherwise} \end{cases}$$

Model :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

REGRESSION COEFFICIENTS

The Model is :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Mean SBP for **Whites** :

$$E(Y | x_1 = 1, x_2 = 0) = \beta_0 + \beta_1$$

Mean SBP for **Blacks** :

$$E(Y | x_1 = 0, x_2 = 1) = \beta_0 + \beta_2$$

Mean SBP for **Others** :

$$E(Y | x_1 = 0, x_2 = 0) = \beta_0$$

Regression Coefficients :

$$E(Y | Whites) - E(Y | Others) = \beta_1$$

$$E(Y | Blackss) - E(Y | Others) = \beta_2$$

Independent Variables X_1 and X_2 :

$$\mathbf{E}(Y \mid \mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2$$

$$\mathbf{E}(Y \mid \mathbf{x}_1 = 0, \mathbf{x}_2 = 0; \text{Others}) = \mu_3 = \beta_0$$

$$\mathbf{E}(Y \mid \mathbf{x}_1 = 1, \mathbf{x}_2 = 0; \text{Whites}) = \mu_1 = \beta_0 + \beta_1$$

$$\mathbf{E}(Y \mid \mathbf{x}_1 = 0, \mathbf{x}_2 = 1; \text{Blackss}) = \mu_2 = \beta_0 + \beta_2$$

Same Null Hypothesis :

$$(\mu_1 = \mu_2 = \mu_3) \Leftrightarrow (\beta_1 = \beta_2 = 0)$$

EQUIVALENCY

- **Same Assumptions**
- **Same Null Hypothesis**
- **In order to prove that they are the same F-test; all we need to show is, say, $SSR=SSB$.**
- **The proof is a bit tedious but not too hard; easier even for the case of the t-test.**

Conclusion:

To compare several means from a completely randomized design, we can do regression with a categorical independent variable – using $(k-1)$ “dummy/indicator variables”. The results are equivalent to those of the one-way Analysis of Variance (one-way ANOVA).

USE OF ALLOCATED CODES

Instead of using two or several dummy variables, some wrongly employ “allocated codes”. For example, a survey on certain frequency (of smoking or use of certain product) responders are asked to choose: Never, Occasional, and Frequent; one single X is then employed with values: $0 = \text{Never}$, $1 = \text{Occasional}$, and $2 = \text{Frequent}$.

The **basic difficulty** with allocated codes is that they define “**metric**” for the categories of the qualitative variable that may or may not be reasonable (here, for example, the difference between Frequent and Occasional is equal to the difference between Occasional and Never).

Instead of a two-sample t-test, we could define a binary covariate and perform a (simple linear) regression.

Instead of an One-way ANOVA (to compare k means), we could define $(k-1)$ dummy/indicator variables and perform a (multiple linear) regression.

What about ANCOVA? What is it and how does it relate to ANOVA and to Regression?

A NEWER POSSIBLE PROBLEM:

Randomization is very crucial because it helps to balanced out the groups. However, even with randomization, groups or arms of a one-factor experiment design might still be unbalanced with respect to some confounder or confounders; without proper adjustment, results might be misleading.

Consider, for example, the case of a clinical trial to compare 3 treatments (Placebo, 300 mg, and 600 mg – 2 doses of a new drug) to lower blood pressure; the major outcome is the SBP *reduction*. A possibility is that **patients in the three groups may have different SBPs at the baseline (prior to intervention) even though patients *were randomized* to treatment groups.**

A similar problem: the comparison of mean SBP and the possibility that the three race groups may have different age distributions.

We can investigate binary covariates (t-test), we can investigate categorical covariates (One-way ANOVA), and – of course - we can investigate continuous covariates.

Of course, a binary covariate or a categorical covariate can be used in the same model with one or more continuous covariates; the “combination” forms an interesting case – that’s ANCOVA.

Example: SBP Versus AGE

Age (X_1)	SBP (Y)
42	130
46	115
42	148
71	100
80	156
74	162
70	151
80	156
85	162
72	158
64	155
81	160
41	125
61	150
75	165

Recall the Example on the left where we have measurements on Systolic Blood Pressure (Y, mm of Hg) and AGE (Years) of **15 women**.

Suppose we have the same data for another group of **10 men**; then we can pool together to form **a sample of size $n=25$** with the same dependent variable Y and 2 covariates:

$$X_1 = \text{Age}$$

$$X_2 = 1 \text{ for women, } 0 \text{ for men}$$

The indicator variable X_2 is also called a “dummy variable”, a “binary indicator”.

THE CASE OF TWO GROUPS

The Model is :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Mean SBP for Women :

$$E(Y | x_2 = 1) = \beta_0 + \beta_1 x_1 + \beta_2$$

Mean SBP for Men :

$$E(Y | x_2 = 0) = \beta_0 + \beta_1 x_1$$

$$E(Y | Women) - E(Y | Men) = \beta_2$$

The model is represented by 2 parallel straight lines; the **difference of the intercepts, β_2 , measures the difference of means SBP between the two genders provided that they are of the same age (for any given age).**

SEVERAL GROUPS

- We can represent “Race”, for example, (with 3 “levels”: White, Black, Others) by defining 2 “Indicator Variables: X_1 (=1 for Whites, =0 for Blacks, =0 for Others) and X_2 (=0 for Whites, =1 for Blacks, =0 for Others): here “Others” is chosen as the “reference level. The slope of X_1 compares Whites to Others and the slope of X_2 compares Blacks to Others.
- We need $(k-1)$ indicator variables to represent a categorical covariate with k categories.

Example: SBP Versus AGE for RACES

$$Y = SBP$$

$$X_1 = Age$$

$$X_2 = \begin{cases} 1 & \text{if White} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if Black} \\ 0 & \text{otherwise} \end{cases}$$

Model :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

REGRESSION COEFFICIENTS

The Model is :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Mean SBP for **Whites** :

$$E(Y | x_2 = 1, x_3 = 0) = \beta_0 + \beta_1 x_1 + \beta_2$$

Mean SBP for **Blacks** :

$$E(Y | x_2 = 0, x_3 = 1) = \beta_0 + \beta_1 x_1 + \beta_3$$

Mean SBP for **Others** :

$$E(Y | x_2 = 0, x_3 = 0) = \beta_0 + \beta_1 x_1$$

Regression Coefficients :

$$E(Y | Whites) - E(Y | Others) = \beta_2$$

$$E(Y | Blackss) - E(Y | Others) = \beta_3$$

The model is represented by **3 parallel straight lines; differences of the intercepts (or vertical distances):**

- (1) β_2 measures the difference of means SBP **between Whites and Others** (provided that they are of the same age),
- (2) β_3 measures the difference of means SBP **between Blacks and Others** (provided that they are of the same age).

The previous few examples are target of “The Analysis of Covariance” (or ANCOVA). This method is used to compare population means adjusted for possible effects of one or several confounders.

The major tool is Multiple Regression

The Analysis of Covariance (ANCOVA) serves the very same main purpose as ANOVA, that is to compares averages or means from different treatments, but it combines the ANOVA method with the Regression method in doing so. The term “ANCOVA” often refers to the Multiple Regression Model without interact terms in which binary/categorical covariate (representing groups) and continuous covariates (confounders) used together.

Appropriate application of ANCOVA tests as part of the statistical analysis of clinical trials provides a more accurate estimate of the real difference among groups; **ANOVA compares the means, ANCOVA compares adjusted means** – mean of study groups adjusted for possible effects of confounders.

Example: SBP Versus AGE for RACES

$$Y = SBP$$

$$X_1 = Age$$

$$X_2 = \begin{cases} 1 & \text{if White} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if Black} \\ 0 & \text{otherwise} \end{cases}$$

Model :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

HOW to compare averages of SBP of three race groups adjusted for Age?

TECHNICAL NOTE:

We can investigate binary covariates, we can investigate categorical covariates, and we can investigate continuous covariates. However, when we have a combination the predictor variables having substantially different magnitudes, we may have serious problems. The inversion of the matrix $X'X$ might run into serious rounding-off errors.

To avoid this computational problem, most packaged computer programs use “Correlation Transformation” to fit Standardized Regression Models then converting the results back to the original measurement scale.

Readings & Exercises

- Readings: A scan through the text's Chapter 8 and sections 7.5-7.6 (pp.294-334) would help.
- Exercises: The following exercises are good for practice, all from chapter 8 of text: 8.6 and 8.21
- Due as Homework: None