



KTH Teknikvetenskap

A very brief introduction to the analysis of experimental data

Harald Lang

Experimental design

In my booklet on econometrics, we analysed *observational* data. In these notes, we are looking at *experimental* data, i.e., the values of the covariates in a regression are set in advance of observing the dependent variable. Most of the problems we may encounter with observational data – sample selection bias, endogeneity, multicollinearity – are avoided when we deal with experimental data. Instead some thought must be put forth in order that we choose a good *experimental design*.

Some jargon

The procedure of selecting the values of the covariates, set in advance, is called *experimental design*.

The process of collecting sample data is called an *experiment*, and the dependent variable to be measured, y , the *response*. The object upon which the response y is measured is called an *experimental unit*. The independent variables, i.e., the covariates, are called *factors*, and their values, set at the design of the experiment, are called *levels*. A *treatment* is a combination of factors levels applied to an experimental unit.

For example (from wikipedia,) in 1747, while serving as surgeon on HMS Salisbury, James Lind carried out a controlled experiment to develop a cure for scurvy.

Lind selected 12 men from the ship, all suffering from scurvy. Lind limited his subjects to men who “were as similar as I could have them”, that is he provided strict entry requirements to reduce extraneous variation. He divided them into six pairs, giving each pair different supplements to their basic diet for two weeks. The treatments were all remedies that had been proposed:

- A quart of cider every day
- Twenty five gutts (drops) of elixir vitriol (sulphuric acid) three times a day upon an empty stomach
- One half-pint of seawater every day
- A mixture of garlic, mustard, and horseradish in a lump the size of a nutmeg
- Two spoonfuls of vinegar three times a day
- Two oranges and one lemon every day.

There is one *factor* (supplement to diet) at six *levels*. The choice of these levels is the *design* of

the *experiment*, the effect on the *experimental unit* (a man) of the various *treatments* is the *response*.

Noise reducing designs

Say for instance that we want to compare the efficiency on learning to drive a car between two different simulators. Let us say that we let twelve randomly chosen individuals (“*experimental units*”) practise on simulator A (“*treatment*” A) and twelve randomly chosen individuals practise on simulator B until they are skilled enough to be eligible for a driving license. The *response*, i.e., the number of ours of practising, will probably vary a lot between individuals practising on *the same* simulator (this is the *experimental error*.)

If we run a regression of hours of practising on an intercept and a dummy for “simulator B”, such that the coefficient for the dummy indicates how much longer those who used simulator B needed to practise compared to those who practised on simulator A, then the variation of hours of those practising on *the same* simulator will show up as the residual in that regression. The standard error of the regression (the estimated standard deviation of the residual) will be a factor in the standard error of the coefficient for the dummy (see relevant formulas in the econometrics booklet.) Hence, we would like to keep this variation as small as possible, in order to be able to make any inference about the sign of that coefficient, i.e., about the relative efficiency of the two simulators.

Some of the variation in performance might depend on differing ages of the individuals, since the capacity to learn to drive varies with age. One way to reduce the variation would thus be to chose individuals of about the same age, but on the other hand, we might want to assess the efficiency of the simulators on individuals with varying ages.

One way to do this is to employ a *randomised block design*.

For instance, we may pick four young individuals, four middle aged and four elderly individuals, and then *randomly* assign two young individuals to each of simulator A and B, and similarly two middle aged and two elderly individuals to each of the two simulators. Each simulator is thus assigned two young, two middle aged and two elderly individuals.

Now we run a regression practising hours on an intercept, a dummy for “middle aged”, a dummy for “elderly” and a dummy for “simulator B”. Some of the variance of the previous residual has been captured by the age dummies, hence “explained” by the model. On the other hand, we have lost two degrees of freedom for the residual (we had ten in the former and have eight in the current model), so it is not certain that the standard error has gone down. But if it has, we have gained some precision in our inference.

Another way to reduce noise is to let more individuals practise on the simulators. This will reduce the standard errors of the coefficients, for these are approximately inversely proportional to the square root of the number of observations. However, this will also increase the cost of the experiment.

There are several experimental designs to consider in order to extract the relevant information in an efficient way, such as “Latin square designs” and “incomplete block designs”. I refer to the literature for these.

One-way ANOVA

Assume that we measure the contamination of calcium (average parts per million) at three locations, A, B and C along the Mississippi River (example taken from [1].) We take several measurements at each location:

A: 42, 37, 41, 39, 43, 41
 B: 37, 40, 39, 38, 41
 C: 32, 28, 34, 32, 30, 33

The measurements are modelled to be observations of normally distributed random variables, $N(\mu_A, \sigma)$, $N(\mu_B, \sigma)$, $N(\mu_C, \sigma)$.

The variance σ^2 is due to measurement errors and other chance, and for the moment we assume they are the same across locations. Now we want to test the hypothesis $H_0: \mu_A = \mu_B = \mu_C$ against the opposite H_1 : “at least two expected values differ.” We can do this by simply running a regression on dummies:

	A	B	C
42	1	0	0
37	1	0	0
41	1	0	0
39	1	0	0
43	1	0	0
41	1	0	0
37	0	1	0
40	0	1	0
39	0	1	0
38	0	1	0
41	0	1	0
32	0	0	1
28	0	0	1
34	0	0	1
32	0	0	1
30	0	0	1
33	0	0	1

We first run a regression of the measurements on dummies for location B and C, i.e., we take location A as reference point.

This is the display from that regression using LINEST in EXCEL.

C	B	const
-9	-1.5	40.5
1.16496	1.22183	0.82375
0.82819	2.01778	
33.74303	14	
274.76471	57	

The estimated difference $\mu_B - \mu_A$ is thus -1.5 , the estimated difference $\mu_C - \mu_A$ is -9 . The F -statistic for the hypothesis that both these differences are equal to zero is 33.74303, and we can compute the p -value to be

$$F_{\text{dist}}(33.74303; 2; 14) = 4.42 \cdot 10^{-6}.$$

An ANOVA analysis (analysis of variance) is typically represented in an ANOVA table:

Source	df	SS	MSS	F	p
location	2	274.765	137.382	33.743	4.42E-006
residual	14	57	4.07143		
total	16	331.765			

The table is interpreted like this: The variation in measurements is defined as the sum of squares $\sum (y_i - \bar{y})^2$ where y_i is the various measurements, and \bar{y} their mean value. This variation is decomposed into two sources: 1) the differing locations, and 2) other sources (residual.)

We find from the result of the regression that total variation of the y -values, $\sum (y_i - \bar{y})^2 = 331.765$, is equal to the sum of squares of the predicted y -values \hat{y}_i from the regression, $\sum (\hat{y}_i - \bar{y})^2 = 274.765$, plus the sum of squared residuals $\sum \hat{\epsilon}_i^2 = 57.0$. The degrees of freedom is 2 for the locations (number of location dummies) and 14 for the residuals (number of observations – number of estimated coefficients, $17-3=14$.) We find these numbers directly from the output from the regression; the “totals” are of course just the corresponding sums.

The mean sum of squares (MSS) is just the sum of squares divided by the degrees of freedom: $137.382 = 274.765/2$, etc. The F -value can then be computed as the ratio $137.382/4.07143 = 33.743$. Finally, the p -value is the probability $\Pr(X \geq 33.743)$ for a random variable whose distribution is the F -distribution with (2, 14) degrees of freedom. In the current case, we of course reject the hypothesis H_0 in favour for H_1 , i.e., the expected values are not all equal. If we had not rejected H_0 , the analysis would have ended here.

Estimating contrasts

Since location A is the reference point in the regression, the coefficients for locations B and C are the estimated differences $\mu_B - \mu_A$ and $\mu_C - \mu_A$. In order to get the estimated difference $\mu_B - \mu_C$ and its standard error, the easiest way is to run the regression with dummies for A and B , i.e., we use location C as reference point:

	B	A	const
	7.5	9	31.5
	1.22183	1.16496	0.82375
	0.82819	2.01778	
	33.74303	14	
	274.76471	57	

so the estimated difference is $\mu_B - \mu_C = 7.5$ with standard deviation 1.22183.

These differences are examples of *contrasts*, i.e., linear combination of the type $c_A\mu_A + c_B\mu_B + c_C\mu_C$ where $c_A + c_B + c_C = 0$. The contrasts of interest are most often simple differences.

Multiple comparisons

Since we accept the hypothesis that *not* all μ_k 's are equal, it is natural to investigate which pairs that differ. We can easily determine the p -values for each of the hypotheses

$\mu_B - \mu_A = 0$, $\mu_C - \mu_A = 0$ and $\mu_B - \mu_C = 0$. We just look at the t -ratios, estimated values over standard errors. I prefer to look at the square of these ratios which come from F -distributions if the corresponding hypotheses are true. Thus, the p -value for $\mu_B - \mu_A = 0$ is

$$F_{\text{dist}}((-1.5/1.22183)^2; 1; 14) = 0.2398,$$

etc. However, if we reject those hypotheses whose p -values fall below, say, 0.05, then the *experimental* error rate is much more than 0.05; i.e., if all hypotheses are true, the probability that we will reject at least one of them (accept that at least some pair is unequal) could be as much as $3 \cdot 0.05 = 0.15$, obviously. In order to keep the experimental error rate ("type 1 error") at a

certain level α , we should accept that a hypothesis is false only if its p -value is at most $\alpha/3$. This is the *Bonferroni* method to deal with multiple hypotheses, there are other, slightly more sophisticated, methods described in the literature, but we are satisfied with this simple one.

A compilation of the estimated contrasts and the p -values for their being = 0 may be presented in a table:

contrasts	estimate	F	p
$\mu_B - \mu_A$	-1.5	1.5072	0.2398
$\mu_C - \mu_A$	-9	59.685	2.05E-006
$\mu_B - \mu_C$	7.5	37.679	2.57E-005

We reject the last two hypotheses, and accept that $\mu_B > \mu_C$ and $\mu_A > \mu_C$. Since the largest of the corresponding p -values is $2.57 \cdot 10^{-5}$, the joint p -value is at most $3 \cdot 2.57 \cdot 10^{-5} = 7.71 \cdot 10^{-5}$. Indeed, if we decide in advance that we should reject any hypothesis whose p -value is at most $2.57 \cdot 10^{-5}$, then the experiment-wise error rate is (at most) $7.71 \cdot 10^{-5}$.

Simultaneous confidence intervals

If we want to compute confidence intervals for the three contrasts and maintain an experiment-wise error rate of, say, 0.05 we should do this with an individual error rate of $0.05/3$, according to Bonferroni's method. This can be accomplished as follows:

$$\text{contrast} = \text{estimated value} \pm s \sqrt{F_{0.05/3}}$$

where s is the estimated standard deviation of the contrast and $F_{0.05/3}$ is the 0.05/3 quantile of the $F(1, 14)$ -distribution, i.e., in EXCEL

$$F_{0.05/3} = \text{FInv}(0.05/3; 1; 14).$$

In the present case:

$$\mu_B - \mu_A = -1.5 \pm 3.32$$

$$\mu_C - \mu_A = -9 \pm 3.17$$

$$\mu_B - \mu_C = 7.5 \pm 3.32.$$

Two-way ANOVA

Assume that we want to compare the strength of two types of glue, A and B, on three materials, Wood, Glass and Rubber. We first consider an additive model: let y be the strength when a glue is applied to a material, and assume that

$$y = \alpha + x_B\beta + z_G\gamma_G + z_R\gamma_R + e.$$

Here x_B is a dummy for glue B, z_G and z_R dummies for material glass and rubber, and e is a random residual which we assume is $N(0, \sigma)$.

The benchmark case is thus glue A on wood. If we have experimental data on all combinations of glue and material, we can of course run the regression and come up with estimates of the coefficients β , γ_G and γ_R . We can also in the usual manner compute the p -values for the hypotheses $H_{0, \text{glue}}: \beta = 0$ (i.e., the two glues give the same strength on average) and $H_{0, \text{material}}: \gamma_G = \lambda_R = 0$ (i.e., the strength is the same for all materials.) We do these tests with the F -test as described in my booklet on econometrics. We may present the analysis in an ANOVA table like this:

source	df	SS	MSS	F	p
material	2	$ \hat{e}_m ^2 - \hat{e} ^2$	SS/df	MSS / MSS(r)	fdist(F; df; df(r))
glue	1	$ \hat{e}_g ^2 - \hat{e} ^2$	SS/df	MSS / MSS(r)	fdist(F; df; df(r))
error	$n-4$	$ \hat{e} ^2$	SS/df		

Here the notation is that of the econometrics booklet; n is the total number of data, $|\hat{e}|^2$ is the residual sum of squares of the full model, $|\hat{e}_m|^2$ and $|\hat{e}_g|^2$ are the residual sum of squares when the material and glue dummies, respectively, are *absent*. MSS(r) is the MSS of the residual, df(r) the df of the residual. The contents of the cells should now be self explanatory.

The reason why we don't have a row "total" is that, in general, the SS of the dependent variable is *not* equal to the sum of the SS:s of material, glue and error. They are equal if we have an equal number of replications (equal

number of data for all combinations of glue and material) but not otherwise.

Assume that we find that the p -value for $H_{0, \text{material}}$ is quite big, such that we contemplate to reduce the model to just have *glue* as explanatory covariates. In order to decide whether we should do this or not, employing the BIC criterion might be of help, but ultimately it is a subjective decision. In any case, it is not advisable to reduce the model unless the standard error of the regression goes down. This is equivalent to the F -value be less than one.

If we decide to reduce the model, we re-estimate the model with the *material* covariates left out. If we have an equal number of replications, the estimated glue coefficients will be exactly the same, but the errors of these estimates should reduce slightly. (If they go up, it means that the regression standard error has increased, so it was presumably a bad idea to reduce the model.)

If we want to make inferences about pair-wise coefficients (difference between *glass* and *rubber*, for instance,) then we proceed just as in the one-way ANOVA case.

Interactions

The previous model is an *additive* one; the difference between the mean strengths of glue A and B is constant across materials. A

model that gives each combination of glue and material a unique mean value is one where *interaction effects* are incorporated:

$$y = \alpha + x_B\beta + z_G\gamma_G + z_R\gamma_R + x_Bz_G\delta_{BG} + x_Bz_R\delta_{BR} + e.$$

Here the product x_Bz_R is a dummy for "glue B on rubber" etc., the benchmark is still "glue A on wood". The ANOVA table now looks like this:

source	df	SS	MSS	F	p
material	2	$ \hat{e}_m ^2 - \hat{e} ^2$	SS/df	MSS/MSS(r)	fdist(F; df; df(r))
glue	1	$ \hat{e}_g ^2 - \hat{e} ^2$	SS/df	MSS/MSS(r)	fdist(F; df; df(r))
interaction	2	$ \hat{e}_i ^2 - \hat{e} ^2$	SS/df	MSS/MSS(r)	fdist(F; df; df(r))
error	$n-6$	$ \hat{e} ^2$	SS/df		

where $|\hat{\epsilon}_i|$ is the residual sum of squares when the interaction dummies are *absent*.

The natural thing to do now is to test if the interaction coefficients can be regarded to be zero. The procedure is the same as above. If we decide to remove the interaction terms, we re-estimate the model as an additive one. If we keep the full model, the mean strength of each combination is given a unique value:

$$\begin{array}{ll} \mu_{AW} = \alpha & \mu_{BW} = \alpha + \beta \\ \mu_{AG} = \alpha + \gamma_G & \mu_{BG} = \alpha + \gamma_G + \beta + \delta_{BG} \\ \mu_{AR} = \alpha + \gamma_R & \mu_{BR} = \alpha + \gamma_R + \beta + \delta_{BR} \end{array}$$

In this case, in order to estimate differences such as $\mu_{BR} - \mu_{AW}$, we re-estimate the model as a regression of y on dummies for the various combinations with – in the current example – glue A on wood as benchmark case.

Problems and alternatives

One-way ANOVA

Let us go back to the one-way ANOVA. The assumption is that data come from normally distributed random variables, with equal variances: $N(\mu_A, \sigma)$, $N(\mu_B, \sigma)$, $N(\mu_C, \sigma)$. Problems arise if either the distributions are not normal, or if variances differ (“heteroskedasticity”).

Heteroskedasticity

Let us start with the second problem; assume that data come from $N(\mu_A, \sigma_A)$, $N(\mu_B, \sigma_B)$, $N(\mu_C, \sigma_C)$. Now we have two more parameters to estimate. Assume that we are primarily interested in the differences $\mu_B - \mu_A$, $\mu_C - \mu_A$ and $\mu_B - \mu_C$. The recommended procedure is as follows: First we compute the obvious point estimates of the three means:

$$\hat{\mu}_A = \frac{1}{6}(42 + 37 + 41 + 39 + 43 + 41) = 40.50$$

$$\hat{\mu}_B = \frac{1}{5}(37 + 40 + 39 + 38 + 41) = 39.00$$

$$\hat{\mu}_C = \frac{1}{6}(32 + 28 + 34 + 32 + 30 + 33) = 31.50$$

and their variances:

$$\hat{v}_A = \frac{1}{6 \cdot (6-1)} \left((42 - \hat{\mu}_A)^2 + \dots + (41 - \hat{\mu}_A)^2 \right) = 0.7833$$

and similarly

$$\hat{v}_B = 0.5 \quad \hat{v}_C = 0.7833$$

Next we compute the (approximate) F -statistic

$$F = \frac{(\hat{\mu}_B - \hat{\mu}_A)^2}{\hat{v}_B + \hat{v}_A} = 1.753$$

Under the null hypothesis that $\mu_B - \mu_A = 0$ this is an observation from an F -distribution with 1 numerator df, and f denominator df, where f is computed by Welch’s formula

$$f = \frac{(\hat{v}_A + \hat{v}_B)^2}{\hat{v}_A^2 / (6-1) + \hat{v}_B^2 / (5-1)} = 8.89.$$

(The numbers 6 and 5 are the number of data for A and B.) We see that we lose about five degrees

of freedom, which is mostly due to the fact that the C-data now are not used for the estimate of the variances \hat{v}_A and \hat{v}_B .

The p -value for the hypothesis is thus

$$p = \text{fdist}(1.753; 1; 8.89) = 0.2185.$$

In EXCEL this p -value can be obtained immediately by employing the *ttest*:

$$p = \text{ttest}(\text{range A}; \text{range B}; 2; 3).$$

Here “2” indicates “two-sided test” and “3” that variances (may) differ and Welch’s procedure is employed.

The other two hypotheses are of course analysed by the same token. If we regard the three hypotheses as one composite hypothesis, we should apply e.g. Bonferroni’s approach.

The new table is now:

contrasts	estimate	F	p
$\mu_B - \mu_A$	-1.5	1.753	0.2185
$\mu_C - \mu_A$	-9	51.702	2.96E-005
$\mu_B - \mu_C$	7.5	43.831	1.03E-004

In the presence of heteroskedasticity, one might question if difference in mean values is the relevant measure. Assume for instance that we have two random variables, $X \in N(50, 15)$ and $Y \in N(55, 5)$. Despite the fact that Y has the larger mean value, the probability that $Y \geq 60$ is less than the probability that $X \geq 60$ (about 16% vs. 25%.) Depending on the situation, we should ponder what exactly we want to know. For instance, imagine that we can choose between two “treatments” resulting in some level X or Y of something, and that a small level in general is favourable, but a level above 60 is especially detrimental. Then we might prefer a treatment resulting in Y rather than X , even though the latter has a lower expected value than the former.

Let us change the question to the following:

If we take a random sample from location A and a random sample from location B, what is the probability that the contamination of sample B is less than that of sample A? In particular, can we reject the hypothesis that this probability is 0.5?

Let us with an abuse of notation write $\text{Pr}(B < A)$ for the above probability. Note that if both the A-data and the B-data come from normal

distributions (possibly with differing variances,) then $\Pr(B < A) = 0.5$ is equivalent to $\mu_B = \mu_A$ (proof left as an exercise.) However, if they come from skewed distributions, these properties are in general not equivalent.

Non normality

A remedy for non normal data that has been proposed in the literature ([2] and references therein) is to transform the original data (the “scores”) to ranks. The transformed data are claimed to be more closely normally distributed.

Note that the property $\Pr(B < A) = 0.5$ is invariant under monotonic transformations, so testing for this property on the ranks is equivalent to testing for it on scores. Since the ranks can be regarded as (close to) normally distributed, we can test for $\Pr(B < A) = 0.5$ by testing for equal means on the ranks. In our Mississippi example the transformed data for A and B are

A: 1.5, 4.5, 8, 8, 10, 11

B: 1.5, 3, 4.5, 6, 8

(ties are given their average ranks.) Welch’s method on these data gives the following p -value:

$$\mu_{\tilde{B}} = \mu_{\tilde{A}}: p = 0.1977$$

where the \sim indicates that the means refer to ranks. Hence, equivalently

$$\Pr(B < A) = 0.5: p = 0.1977.$$

etc. Note that when we compare location A against location B, the data from location C do not enter the calculations.

Point estimate for $\Pr(B < A)$

If we want a point estimate of the probability $\Pr(B < A)$ we proceed as follows. We note the number of pairs (B,A) where $B < A$, giving ties the value 0.5. We divide that number with the product (number of A-data)·(number of B-data). In our case, there are 20 pairs where $B < A$ and four ties, so we get $22/(6 \cdot 5) = 22/30$. Hence, our point estimate is

$$\Pr(B < A) = 0.7333.$$

What we have actually estimated is $\Pr(B < A) + 0.5 \Pr(B = A)$, but if the original data come from continuous distributions, then $\Pr(B = A) = 0$.

Two-way ANOVA

The problems in the one-way ANOVA case may of course also appear in a two-way ANOVA analysis. The remedies should be the same, but in practise it may not work, since without the restriction that all variances are the same, there is an excessive number of parameters to estimate.

Another issue with a two-way ANOVA analysis is the number of data required in order to test for interactions. If we have precisely one observation for each combination (“in each cell”), then in the regression, all error terms will come out zero. Indeed, the number of parameters to be estimated is exactly equal to the number of observations. Hence, in order to get an estimate of the standard error, we need at least one more observation. Preferably, we should have the same number of measurements (hence at least two) for each treatment combination, since the problem with heteroskedasticity is least in this case.

Random models

Let us go back to the very first example in the section about one-way ANOVA. But now our concern is the following: “How much, if at all, does the contamination of calcium vary across locations in the Mississippi River?” Phrased differently: If we select a location at random, then the mean contamination there is a random variable. We denote that random variable $\mu + d$ where μ is the expected value and $d \in N(0, \tau)$. If we take one measurement at that location, the observed value is a random variable which we denote

$$y = \mu + d + e,$$

where $e \in N(0, \sigma)$ and we assume that d and e are independent. The distribution of y is thus

$$N(\mu, \sqrt{\tau^2 + \sigma^2}).$$

The data given in the example can now be regarded as observations of y , and our task is to estimate the standard deviation τ and possibly also σ . What has changed as compared to the example as stated in the one-way ANOVA section is that we now think of the three locations A, B and C as *random draws* of locations, not deterministically selected locations.

We have 17 observations on y , but they are obviously *not* independent; indeed, observations within a fixed location have the common component d , so their covariance is τ^2 whereas observations from different locations are independent (covariance = 0.) The entire covariance matrix for the 17 observations is thus a block matrix:

$$\begin{pmatrix} B_A & 0 & 0 \\ 0 & B_B & 0 \\ 0 & 0 & B_C \end{pmatrix}$$

where B_A is the 6×6 -matrix with $\tau^2 + \sigma^2$ in the diagonal, and τ^2 in the off-diagonal entries. Similarly for B_B and B_C . We can now estimate μ , τ and σ by Maximum Likelihood (ML) estimation, and also get the standard errors of these estimates in the usual way for ML-estimation.

In the literature one is often advised a different approach, where we estimate σ^2 and $\tau^2 + \sigma^2/n$ (where n is the number of observations at each location, which is assumed to be the same across locations) from linear regressions, and then get an estimate of τ^2 by taking the appropriate difference. This is computationally much simpler, but also much less accurate. This was a reasonable approach when computational costs were considerable, but today they are negligible, so this approach should not be employed today.

Mixed models

There are situations where some of the covariates are random and some deterministic. They are also to be estimated by ML in a similar manner.

We don't go into further details on random and mixed models here.

References

- [1] Belz M.H., *Statistical Methods for the Process Industries*, Macmillan 1973
- [2] Ruxton G.D., *The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test*, Behavioral Ecology 17(4) 2006