Elements of Regression Analysis

Harald Lang

I. Introduction

A very common situation in statistical analysis is that we want to relate the outcome of a random variable y – the "response variable" – to the values of some other variables $x_1,...,x_k$ – the *covariates*. The covariates "explain" the outcome of the variable y to some extent, and the deviation between the actual outcome y and the explained part $\hat{y}$ is the unexplained "residual" or "disturbance term". The covariates can either be *observational* – outcomes that we do not control – or *experimental* – results of experiments controlled by us. They can either be *quantitative* or *qualitative*. The interpretation of the relation under study can either be *predictive* or *structural.* In all cases, the process of modelling these relations and performing the statistical analyses of the relationships is called *regression analysis.*

Before we embark on the topic on regression analysis, we need some facts on hypothesis testing.

Assume that we want to test if the value of the unknown parameter θ is equal to θ_0. We can think of it this way: we actually consider *all* hypotheses $\theta = a$ for all (feasible) values of a. We know that precisely one of these hypotheses is true, the others are false. If we accept an error rate (probability of a type 1 error; accepting a false alternative hypothesis) α, then we can assume that the data we are using for the test we are going to perform are such that the p- value for the *true* hypothesis comes out greater than α. Indeed, the probability that this p- value comes out at most α is (at most) α and is hence a risk we are accepting to take. *Therefore, any test we perform for a hypothesis $\theta = a$ that comes out with a p- value at most α must be false, and hence can be rejected .*

Let us look at a simple example. Let us say that x is an observation of a normally distributed random variable with unknown mean value θ and variance equal to 1. The null hypothesis is that $\theta = 0$. Assume that we choose $\alpha = 0.01$ and that $x = 2.7$. Under the null, $(2.7 - 0)^2$ is an observation of a $\chi^2(1)$-variable. The p-value for the null is

$\Pr(\chi^2(1) > 7.29) = 0.0069$, so we reject the null. *But by the same token, we will reject any hypothesis* $\theta = a$ *for any* $a \le 0$. As a consequence, we do not only reject that $\theta = 0$, *we accept that* $\theta > 0$. *More specifically, we accept that* $0.124 < \theta < 5.276$. We have *inverted* a hypothesis test to a *confidence interval.*

If we test for more than one parameter value, say θ_1, θ_2 and θ_3, then we assume that the data are such that when we test for the *true* value θ_1° of θ_1, then the *p*-value for that test will come out $> \alpha/3$, and similarly for θ_2 and θ_3. The probability for this is at least $1 - 3 \cdot (\alpha/3) = 1 - \alpha$. This procedure for treating multiple hypotheses is called *Bonferroni's* method. Hence, we reject the test for any of the tree parameters that comes out with a *p*-value $\le \alpha/3$ (if we test for three parameter values.) Confidence intervals are treated similarly; three intervals with confidence level $1 - \alpha/3$ have a joint covering probability of (at least) $1 - \alpha$.

Note the difference between testing *several values for a single parameter* and testing *values for several* (say three) *parameters*. In the first case, there is precisely one true value, in the second there are several (three) true values. Bonferroni's method applies to the second case.

II. The Linear Regression Model

(when everything is almost perfect)

The specification of a linear regression model is

$$y_i = \sum_{j=0}^{k} x_{ij}\beta_j + e_i, \quad i = 1,\ldots,n\,.$$

Here y_i is an observation of the the *dependent* random variable y whose value depends on the *covariates* (or *explanatory variables*) $x_{\bullet j}$ plus an additional random variable, the *residual* e_i. The covariates are regarded as deterministic, i.e. fixed in repeated samples, whereas the residuals are random variables which are assumed to be *independent between observations* and such that

$$\mathrm{E}(e_i) = 0 \;\text{ and }\; \mathrm{E}(e_i^2) = \sigma_i^{\,2} \;\text{ and }\; \mathrm{E}(e_i^4) < \infty$$

where the σ_i:s are unknown. The covariate $x_{\bullet 0}$ is the constant 1, and β_0 is the *intercept*.

If we introduce

$$x_i = (x_{i0} \quad \ldots \quad x_{ik}), \quad i = 1,\ldots,n \quad \text{and}$$
$$\beta = (\beta_0 \quad \ldots \quad \beta_k)^t$$

then the model may be written as

$$y_i = x_i\beta + e_i$$

Sometimes we suppress the observation index i and just write

$$y = x\beta + e$$

This is the *heteroskedastic* version of the linear model, meaning that the variances of the e_i:s are not all the same. Traditionally, in textbooks, one assumes that all σ_i:s are the same, which is often a rather unjustified assumption. This is the *homoskedastic* version of the model. Historically, this was the standard assumption, since it simplifies both the theory and the computations. Nowadays, when

computation is fast and cheap, applied statisticians typically assume heteroskedasticity since this generalisation comes at no cost. Unfortunately, many computer software (such as EXCEL) still use the homoskedastic assumption as the default model (EXCEL has currently no other option.)

It is convenient to employ matrix notation:

$$Y = X\beta + e \text{ where } E(e) = 0 \text{ and } E(ee^t) = D(\sigma_i^2)$$

where now Y is an $n \times 1$-matrix of random variables, X is an $n \times (k+1)$-matrix and e is an $n \times 1$-matrix of random variables. $D(\sigma_i^2)$ is a diagonal matrix whose diagonal elements are σ_i^2, $i = 1 \ldots n$.

The parameters β_j are unknown and are to be estimated from data. The use of the model can either be for *prediction,* or may be given a *structural* interpretation which allows for hypotheses testing.

Here is an example of a structural interpretation. Assume we want to assess if females, *ceteris paribus,* get lower salaries than males, as is often claimed. We can then estimate the linear model

$$y = \beta_0 + x_1\beta_1 + \cdots + x_k\beta_k + e$$

where y is log(wage), x_1 is an indicator variable (typically called a *dummy*) for female (i.e., $x_1 = 1$ for females and $x_1 = 0$ for males,) and the other covariates are age, years_of_education, years_of_working_ experience, etc.; i.e., characteristics that we believe influence the wage. Females' wages are then on average $\exp(\beta_1)$ times that for males, and we interpret a negative value of β_1 as a confirmation of the claim that females on average get lower wages than males, *ceteris paribus.*

A *structural interpretation* means that we consider the *covariates to influence the dependent variable,* but not the other way round. This need not be the case for a *prediction.* For example, assume that we

want to predict a student's performance (grading) on a statistics course, and use his previous grades on mathematics courses he has taken as covariates. Then, of course, these grades have no *influence* on his performance on the statistics course, but they have some predictive power for obvious reasons: first, they measure some kind of ability (a "hidden" characteristic) that is also useful in a statistics course and, secondly, it measures his mathematical knowledge, which is also useful in a statistics course.

Another example is this: assume we want to assess the engine power of a car, and use the time it takes for it to accelerate from 0 kmph to 50 kmph and its weight as covariates. Of course, neither the time it takes to accelerate the car, nor its weight, *causes* the engine to have a certain power. Here the impact goes the other way: the power influences the time it takes to accelerate.

In the first example – women's wages as compared to men's – it is important that we include all covariates we believe influence a person's wage. Otherwise it could be that women on average have less working experience than men, for example, and this causes them to get lower wages, whereas gender has no impact on wages. If we want to assess if gender, *ceteris paribus,* has an impact on wages, then we have to be careful with the modelling. If we just want to make a *prediction* of a person's wage, then we need not worry about if the predictive power of "gender" depends on gender itself, or if it is just correlates with, say, working experience or education, or whatever. Most of the problems we will talk about later concern structural interpretations; it is generally much easier to model and estimate equations for prediction.

Estimation

The *OLS* estimate (Ordinary Least Squares) of β, $\hat{\beta}$, is the value $\hat{\beta}$ of β that minimises the sum of the squares $\hat{e}^t\hat{e} = |\hat{e}|^2$ of the *residuals* $\hat{e} = Y - X\hat{\beta}$. This is achieved by solving the *normal equations*

$$X^t\hat{e} = 0$$

for $\hat{\beta}$. Indeed, let $\tilde{\beta}$ be any other estimate of β, and define $\tilde{e} = Y - X\tilde{\beta}$. Let $\Delta\beta = \hat{\beta} - \tilde{\beta}$. Now

$$\tilde{e} = Y - X\tilde{\beta} = X\hat{\beta} + \hat{e} - X\tilde{\beta} = X\Delta\beta + \hat{e}$$

But $X\Delta\beta$ and $\hat{e}$ are orthogonal (the normal equations), hence, by "Pythagoras' theorem",

$$|\tilde{e}|^2 = |X\Delta\beta|^2 + |\hat{e}|^2 \geq |\hat{e}|^2$$

Q.E.D.

From the normal equations it follows that the OLS estimate of β is

$$\hat{\beta} = (X^t X)^{-1} X^t Y .$$

Note that $\hat{\beta} = (X^t X)^{-1} X^t (X\beta + e) = \beta + (X^t X)^{-1} X^t e$, or

$$\hat{\beta} = \beta + (X^t X)^{-1} X^t e$$

and hence $\hat{\beta}$ is an unbiased estimate of β; $\mathrm{E}(\hat{\beta}) = \beta$.

The covariance matrix for $\hat{\beta}$ can now be computed:

$$\mathrm{Cov}(\hat{\beta}) = \mathrm{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^t]$$

$$= (X^t X)^{-1} X^t D(\sigma_i^2) X (X^t X)^{-1}.$$

A consistent estimate of this covariance matrix is

$$\hat{\mathrm{Cov}}(\hat{\beta}) = \frac{n}{n-k-1} (X^t X)^{-1} X^t D(\hat{e}_i^2) X (X^t X)^{-1}. \qquad \text{(ii.1)}$$

The factor $n/(n-k-1)$ needs an explanation. Since we minimise the sum of squares $|\hat{e}|^2$, we will underestimate the true value $|e|^2$. The factor $n/(n-k-1)$ is an *ad hoc* compensation for this.

The BLUES

For the homoskedastic model, the OLS estimator is the Best Linear Unbiased EStimator (BLUES). By "linear" is meant an estimator of the type $\tilde{\beta} = AY$. We get $\tilde{\beta} = AY = AX\beta + Ae,$ so in order that $\tilde{\beta}$ be unbiased, whatever value of $\beta,$ we must require that $AX = I.$ The covariance matrix for this estimator is $AA^t\sigma^2$ whereas the homoskedastic OLS covariance matrix is $(X^tX)^{-1}\sigma^2,$ as can be seen as above. We now prove that

$$AA^t - (X^tX)^{-1} \geq 0,$$

meaning that this matrix is positive definite or semi definite. This then proves that OLS is the BLUES, in the sense that it minimises the variances of the individual β estimates, and every linear combination of these.

Define $A_0 = (X^tX)^{-1}X^t.$ Note that, since $AX = I,$

$$AA_0^t = A_0A^t = (X^tX)^{-1}.$$

Hence

$$\begin{aligned} 0 &\leq (A - A_0)(A - A_0)^t \\ &= AA^t - AA_0^t - A_0A^t + A_0A_0^t \\ &= AA^t - (X^tX)^{-1} \end{aligned}$$

Q.E.D.

However, for the heteroskedastic model, OLS is inefficient. It is therefore commendable to formulate the regression model in such a way that it be as close to homoskedastic as possible.

There are estimators that are more efficient, asymptotically and theoretically, than OLS for the heteroskedastic model, but these are not as robust as OLS, so in practise, OLS is the estimator of choice.

Tests and Confidence Intervals for the β:s

The standard error $\mathrm{SE}(\hat{\beta}_j)$ (the estimate of the standard deviation of the estimator $\hat{\beta}_j$) is the square root of the corresponding diagonal element of $\hat{\mathrm{Cov}}(\hat{\beta})$. This standard error is reported by EXCEL (for the homoskedastic model) and other programmes that can perform linear regression analysis. A confidence interval at level $1-\alpha$ for β_j is

$$\hat{\beta}_j \pm \sqrt{F_\alpha(1, n-k-1)}\,\mathrm{SE}(\hat{\beta}_j) \qquad \text{(ii.2)}$$

where $F_\alpha(1, n-k-1)$ is the α quantile of the F-distribution with one numerator degrees of freedom and $n-k-1$ denominator degrees of freedom. An $F(1, n-k-1)$ -statistic for the hypothesis $\beta_j = \beta_j^0$ is

$$F = \left(\frac{\hat{\beta}_j - \beta_j^0}{\mathrm{SE}(\hat{\beta}_j)} \right)^2 .$$

The p-value for the hypothesis is $\Pr(F(1, n-k-1) > F)$.

R^2 and Effect Size

We run a regression of y on some covariates and compute the sum of residuals $|\hat{e}|^2$ and also y on only an intercept, which gives the residual sum of squares $|\hat{e}_*|^2$. The difference between these two sums of squares is the amount of variation "explained" by the covariates. The relative size of this "explained" part is denoted R^2:

$$R^2 = \frac{|\hat{e}_*|^2 - |\hat{e}|^2}{|\hat{e}_*|^2}$$

It is a measure of *goodness of fit* and is called the "coefficient of determination". It is also equal to the square of the (sample) correlation coefficient between y and $x\hat{\beta}$. See the exercises for other interpretations of R^2. There is also an adjusted R^2 (often denoted $\bar{R}^2$) where there is an adjustment for degrees of freedom, such that the adjusted R^2 is somewhat lower than R^2.

This can of course be generalised; indeed, if we run a regression with and without a certain covariate (or set of covariates) then the *effect size* called *partial eta squared* is defined as

$$\eta^2 = \frac{|\hat{e}_*|^2 - |\hat{e}|^2}{|\hat{e}_*|^2},$$

or, equivalently

$$\eta^2 = \frac{R^2 - R_*^2}{1 - R_*^2}$$

where R^2 and R_*^2 are the coefficients of determination for the full and the "restricted" regression; $\hat{e}$ and $\hat{e}_*$ corresponding residuals.

It is convenient to compute η^2 directly from the output of the full regression (i.e., all covariates included.) This is done as follows: Assume that the model is

$$y = x_1\beta_1 + x_2\beta_2 + e$$

(here x_1 and x_2 are row matrices, and β_1 and β_2 column matrices) and we want to compute the effect size of x_2. Let W_2 be the sub-matrix of $(X^tX)^{-1}$ corresponding to the x_2-variables (i.e., the last r rows and columns, if x_2 contains r covariates.) Define

$$q = \hat{\beta}_2^t W_2^{-1} \hat{\beta}_2 .$$

Now

$$q = |\hat{e}_*|^2 - |\hat{e}|^2, \quad \text{so} \quad \eta^2 = \frac{q}{|\hat{e}|^2 + q}. \tag{ii.3}$$

It is a good practise to include η^2 in regression output tables.

Hypothesis Testing concerning Several β Values; the *F*-test

Assume we want to test the null hypothesis that a number r of the β:s are all equal to zero. A test statistic for the hypothesis is

$$F = \frac{1}{r}\hat{\beta}_2^t \hat{V}_2^{-1}\hat{\beta}_2$$

which has an approximate $F(r, n-k-1)$ distribution under the null, and we reject the hypothesis if F is large. The p-value for the hypothesis is $\Pr(F(r, n-k-1) > F)$. The notation is as follows: Assume that the model is

$$y = x_1\beta_1 + x_2\beta_2 + e$$

and H_0 – the hypothesis to be tested – is that $\beta_2 = 0$ (here x_1 and x_2 are row matrices, and β_1 and β_2 column matrices) and β_2 consists of r parameters. $\hat{V}_2$ is the estimated covariance matrix for $\hat{\beta}_2$ (i.e., we keep the last r columns and the r last rows in the estimated covariance matrix for $\hat{\beta}$ and delete the others.)

We can also compute a confidence set for β_2: with confidence level $1-\alpha$, the true parameters β_2 lie in the region (an r-dimensional ellipsoid)

$$(\beta_2 - \hat{\beta}_2)^t \hat{V}_2^{-1}(\beta_2 - \hat{\beta}_2) \le rF_\alpha(r, n-k-1).$$

Other Linear Restrictions

A general linear restriction may be written

$$R\hat{\beta} = \mu$$

where R is an $r \times (k+1)$-matrix and μ an $r \times 1$-matrix. The simplest way to handle this problem is to rewrite the model to convert the problem to the situation where the hypothesis is that a set of coefficients is zero. I show how this is done in an example. Let us say that the hypothesis is that $\beta_1 + \beta_2 = 1$ and $\beta_3 = 2\beta_4$ and that the model is

$$y = \beta_0 + \sum_{j=1}^{5} x_j\beta_j + e$$

Define γ_2 and γ_3 such that $\beta_2 = 1 - \beta_1 + \gamma_2$ and $\beta_3 = 2\beta_4 + \gamma_3$ and

rearrange:

$$\begin{aligned} y - x_2 = \beta_0 &+ (x_1 - x_2)\beta_1 \\ &+ (2x_3 + x_4)\beta_4 + x_5\beta_5 \\ &+ x_2\gamma_2 + x_3\gamma_3 + e \end{aligned}$$

We thus generate a new dependent variable ($y - x_2$) and two new covariates ($x_1 - x_2$) and ($2x_3 + x_4$) and employ the F-test on the hypothesis $\gamma_2 = 0$ and $\gamma_3 = 0$. Note that since we have a new dependent variable, the R^2 statistic from this regression is incompatible with that from the original formulation of the full model.

Prediction

The linear model is often used for prediction. Given the covariates x_0 (a row matrix), the predicted value of the corresponding y, y_p, is

$$y_p = \hat{\beta}_0$$

where $\hat{\beta}_0$ is the estimate from

$$y = \beta_0 + (x - x_0)\beta + e$$

This prediction contains two components of uncertainty: the residual, which is set to zero in the prediction equation, and the fact that the estimated value $\hat{\beta}_0$ rather than the true β_0 is employed. Thus, the error in the prediction is

$$e_p = e_0 + \beta_0 - \hat{\beta}_0$$

whose total variance is

$$\mathrm{Var}(e_p) = \sigma_0^2 + \mathrm{Var}(\hat{\beta}_0)$$

which is estimated to

$$\hat{\mathrm{Var}}(e_p) = \hat{\sigma}_0^2 + \hat{\mathrm{Var}}(\hat{\beta}_0)$$

where $\hat{\sigma}_0^2$ is an estimate of $\sigma_0^2 = \mathrm{Var}(e_0)$. If the model is homoskedastic, an unbiased estimate is

$$\hat{\sigma}_0^2 = \frac{|\hat{e}|^2}{n-k-1}.$$

III. The Logit

In some cases y_i is a probability. For instance, we might want to estimate the probability that an unemployed person gets a job during one month, conditional on covariates like age, education, experience, gender etc. In this case a linear specification

$$y_i = x_i\beta$$

might be inappropriate. For instance, it may happen that for some values of the covariates $x_i\beta > 1$ or < 0. Two common specifications are the *probit* and the *logit* specifications. The probit specification is

$$y_i = \Phi(x_i\beta)$$

where Φ is the cumulative normal density function; the logit specification is

$$y_i = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)} = p(x_i\beta)$$

In practise they differ very little, so we focus on the latter, since it is computationally much simpler.

The data on y_i are given as dummy variables y_i such that y_i is equal to one if the event under study occurred, and zero otherwise. The estimation is then by Maximum Likelihood, i.e., one maximises the log-likelihood function

$$\ln(L) = \sum_{i=1}^{n} \ln[(2y_i - 1)p(x_i\hat{\beta}) + 1 - y_i]$$

over $\hat{\beta}$. For testing, one may use the *likelihood-ratio test:* if $\ln(L_*)$

is the log-likelihood function when the restrictions are imposed, then under the null (i.e., that the restrictions are true)

$$2\ln(L) - 2\ln(L_*)$$

is approximately a $\chi^2(r)$ variable, where r is the number of restrictions. One may also use bootstrapping, (see section "The Bootstrap".)

Unfortunately, the logit (and probit) is sensitive to misspecifications.

With the logit (or probit) specification the interpretation of the β-coefficients are less obvious than in the OLS case. A positive value of a β_j implies that the corresponding covariate has a positive influence on the probability y, but to what degree depends on the value of the other covariates. However, we can rewrite the specification as

$$\ln\left(\frac{y_i}{1-y_i}\right) = x_i\beta$$

so we see that the impact of a covariate on the *log-odds* is measured by the corresponding β-coefficient.

Exercises

1. Let $\hat{e}_i$ be the estimated residuals of a regression of y_i onto some covariates. Show that

$$\sum_1^n y_i\,\hat{e}_i = \sum_1^n \hat{e}_i^2.$$

2. We run a regression

$$y = \hat{\beta}_0 + x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + x_3\hat{\beta}_3 + \hat{e}$$

where $\hat{}$ denotes estimated values. Next we run the regression

$$x_3 = \hat{\alpha}_0 + x_1\hat{\alpha}_1 + x_2\hat{\alpha}_2 + \hat{u}.$$

Prove that $\hat{u}^t\hat{e} = 0$.

3. Assume that you run a regression of y onto two covariates x_1 and x_2:

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + e.$$

a) If the estimated regression coefficient for x_1 is zero when x_2 is regressed on (an intercept and) x_1, then the estimated value of β_1 will be the same as when you leave out x_2 from the regression. Prove that!

b) Prove that this is not true if the coefficient for x_1 in the regression described in a) is different from zero.

c) If the coefficient for x_1 *is* equal to zero, is there any reason to include the x_2 covariate in the regression? (We are only interested in the β_1 coefficient.)

4. When we run a regression $y = \beta_0 + \sum_{j=1}^{k} x_j\beta_j + \text{residual}$ we get a value of R^2, which is defined in the text. Show that R^2 can also be described as

$$R^2 = \frac{\text{Var}(\hat{y})}{\text{Var}(y)},$$

where "Var" is the sample variance, and

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{k} x_j\hat{\beta}_j .$$

Furthermore, show that R^2 is also equal to the square of the sample correlation coefficient between y and $\hat{y}$.

IV. When Everything is Not so Perfect

Multicollinearity

Assume that we run a regression of a variable *y*, say log of wage, on dummies[1] (male) and (female) and an intercept:

$$\log(\text{wage}) = \beta_0 + (\text{male})\beta_1 + (\text{female})\beta_2 + e$$

(so "(male)" is equal to one if the person is a man and zero if it is a woman, etc.) It is easy to see that the OLS estimate does not have a unique solution. Indeed, we can add any number a to β_1 and β_2 and subtract a from β_0 and get the same residuals. The problem is that the intercept (the covariate 1) and the two dummies are linearly dependent. The problem is labelled (*perfect*) *multicollinearity.* This is a model specification error which renders OLS estimation impossible. In the example given here, the remedy is to remove one of the gender dummies. The coefficient for the remaining dummy estimates the extra wage persons of this gender enjoys.

More often the problem is imperfect multicollinearity, often just labelled "multicollinearity". This means that at least one of the covariates is is highly correlated with a linear combination of the other covariates (but not a perfect linear combination of these.)

Say, for instance, that you run a regression of log wage on age, education (in years) and working experience (in years). The problem is that age $\approx$ 6 + education + experience for most persons.

Multicollinearity (i.e., imperfect) *is not a specification error*, it is just a nuisance, since it causes the standard errors of one or more more of the regression coefficients to be very large, and hence the point estimates of the those coefficients to be very imprecise. However, these standard errors are decreasing in *n*, the number of observations, so the problem with multicollinearity is in a way equivalent to few observations. See appendix 5.

---

[1] See the section "Model Selection" below.

Micronumerosity

"*Econometrics texts devote many pages to the problem of multicollinearity in multiple regression, but they say little about the closely analogous problem of small sample size in estimating a univariate mean. Perhaps that imbalance is attributable to the lack of an exotic polysyllabic name for "small sample size." If so, we can remove that impediment by introducing the term micronumerosity.*" (Arthur S. Goldberger, A Course in Econometrics, 1991.)

This is said in jest, of course; "micronumerosity" is not a word. But the problem with few data is a real problem. The classical assumptions on the regression model is that the error terms e_i are independent and follow a normal distribution with mean zero and a common variance σ^2. We have assumed only independence and mean zero, plus the extra condition that $\mathrm{E}(e_i^4) < \infty$. In applications the assumptions on normal distribution and homoskedasticity (see below) are at best approximately appropriate, so instead we rely on asymptotics; more precisely the Central Limit Theorem. It is for this we need the assumption $\mathrm{E}(e_i^4) < \infty,$ but we need also require that the number of observations is not too small.

If the number of observations n is small, then the asymptotics does not kick in, so we need to rely on the assumptions on normally distributed error terms and homoskedasticity. Hence we employ the homoskedastic formulas for standard errors, confidence intervals and tests in these cases. See the paragraph "Why do we assume homoskedasticity in ANOVA?" in section VIII.

Heteroskedasticity

A common assumption is that all residuals e_i have the same standard deviation σ. This assumption is called *homoskedasticity*. The model specification is thus

$$y_i = \sum_{j=0}^{k} x_{ij}\beta_j + e_i, \quad i = 1,\ldots,n$$

where

$$\mathrm{E}(e_i) = 0 \text{ and } \mathrm{E}(e_i^2) = \sigma^2$$

and the e_i:s are independent between observations. When this specification is employed, the computer programme will compute standard errors and F-statistics from a covariance matrix which is estimated as

$$\hat{\mathrm{C}}\mathrm{ov}(\hat{\beta}) = (X^t X)^{-1} s^2, \text{ where} \qquad \text{(iv.1)}$$

$$s^2 = \frac{|\hat{e}|^2}{n-k-1}.$$

If the model has heteroskedastic residuals and is misspecified as homoskedastic, then this causes the standard deviations of the parameter estimates to be inconsistent and renders the F-test invalid – the point estimates of the coefficients are still the same, though. Let us look at an example.

Assume that $y_1,\ldots,y_{2m}$ are observations of a normally distributed variable with unknown expected value μ_1 and unknown variance equal to 1, and that $y_{2m+1},\ldots,y_{3m}$ are observations of a normally distributed variable with unknown expected value μ_2 and unknown variance equal to 2, and and that all observations are independent. We want to test the hypothesis that $\mu_1 = \mu_2$, or compute a confidence interval for the difference $\mu_1 - \mu_2$. We express the situation as a linear regression model:

$$y_i = \beta_0 + x_i\beta_1 + e_i,\ i = 1,\ldots,3m$$

where $x_\bullet$ is a dummy variable for the first $2m$ observations, i.e., $x_i = 1$ for $i = 1,\ldots,2m$ and $x_i = 0$ for $i = 2m+1,\ldots,3m$. The hypothesis is now that $\beta_1 = 0$, and a confidence interval for $\mu_1 - \mu_2$ is the same as a confidence interval for β_1. The true variance for the estimate $\hat{\beta}_1$ is easy to compute, it is equal to $2.5m^{-1}$. If the model is misspecified as a homoskedastic regression model, the asymptotic value of the estimated variance is $2m^{-1}$. Indeed, the estimated variance is computed as

$$\frac{1}{3m-2}\left(\sum_{1}^{2m}(y_i-\overline{y}_{1:2m})^2+\sum_{2m+1}^{3m}(y_i-\overline{y}_{2m+1:3m})^2\right)\left(\frac{1}{2m}+\frac{1}{m}\right)$$

whose expected value is about $2/m$ for large m. In contrast, the heteroskedastic estimated variance is

$$\frac{3m}{3m-2}\left(\frac{1}{4m^2}\sum_{1}^{2m}(y_i-\overline{y}_{1:2m})^2+\frac{1}{m^2}\sum_{2m+1}^{3m}(y_i-\overline{y}_{2m+1:5m})^2\right)$$

whose expected value is about the correct value $2.5/m$ for large m.

Remedies for Heteroskedasticity

Reformulate the Model

The first thing to do is to try to reformulate the model so as to get rid of the heteroskedasticity, see the section "Model Selection". The reason for this is that OLS is more efficient if the residuals are close to homoskedastic.

White's Consistent Variance Estimator

Halbert White wrote an influential article in 1980 where he gave a consistent estimator for the covariance matrix for the heteroskedastic regression model. His covariance estimator is the one in this text without the leading factor $n/(n-k-1)$. The estimator in this text is the default heteroskedasticity robust covariance matrix estimator in the programme Stata.

Most regression computer programmes have the option to use "robust errors" or some such, which means employing some variant of White's estimator. It is advisable to *always* use this feature when available (see however the section "Why do we assume homoskedasticity in ANOVA?")

The Bootstrap

It is a good idea to use *bootstrap* to test hypotheses and estimate confidence intervals in case of for example non-normality of the

residuals, 2SLS (described later) and also in case of logit estimation, since in these cases the standard methods have poor properties in small samples. I will describe one way to do this. Let us say that we want to test if a sub-vector θ of β is equal to 0. Let $\hat{\theta}$ be the point estimate of θ and let $\hat{\theta}_j^*$, $j = 1,\ldots, b$, be the the corresponding estimates from the *bootstrap resamples* (see below). Now estimate a covariance matrix for $\hat{\theta}$ (the accuracy of this estimate is not crucial):

$$\hat{V} = \frac{1}{b}\sum_{j=1}^{b}\left(\hat{\theta}_j^* - \hat{\theta}\right)\left(\hat{\theta}_j^* - \hat{\theta}\right)^t.$$

Next compute the constant c:

$$c = \hat{\theta}^t \hat{V}^{-1} \hat{\theta}$$

and compute the number m of j:s for which

$$\left(\hat{\theta}_j^* - \hat{\theta}\right)^t \hat{V}^{-1}\left(\hat{\theta}_j^* - \hat{\theta}\right) > c$$

The p-value for rejecting the null hypothesis is now m/b.

If θ is a single parameter, we can compute a confidence interval for it as

$$\theta = \hat{\theta} \pm c$$

where c is chosen such that a fraction pb of the $\hat{\theta}_j^*$:s falls outside of the interval $\hat{\theta} \pm c$. The bootstrap resampling can be done in various ways, one of which is "wild bootstrap" which is as follows: Let us say that the data set consists of n data of the form $\{(x_1, y_1),\ldots,(x_n, y_n)\}$, where the x_i:s are row vectors. A resample is a new set of n data created as follows: After the regression is run, keep the residuals $\hat{e}_i$, $i = 1,\ldots, n$ and then create the *bootstrap resample* $\{(x_1, y_1^*),\ldots,(x_n, y_n^*)\}$ such that

$$y_i^* = \begin{cases} y_i & \text{with probability 0.5} \\ y_i - 2\hat{e}_i & \text{with probability 0.5} \end{cases}$$

(i.e., we change the sign of the residual with probability 1/2.) With these data we run the regression again, and obtain new estimated parameters θ_1^*. This is repeated b times (say, 2'000 times,) to give the bootstrap estimates $\theta_j^*,\ j = 1, \ldots, b.$

Employed to logit regression, we resample as follows: the resample is $\left\{(x_i, y_i^*)\right\},\ i = 1,\ldots, n,$ where

$$y_i^* = \begin{cases} 1 & \text{with probability } p(x_i\hat{\beta}) \\ 0 & \text{with probability } 1 - p(x_i\hat{\beta}). \end{cases}$$

V. Model Selection

The use of dummy variables

In many model specifications some covariates are *qualitative* and then *dummy variables* come in handy. Assume, for example, that you want to test if returns to schooling (on wage) is the same for men and women. You can then specify a model like this:

$$\begin{aligned} \ln(\text{wage}) = \beta_0 &+ (\text{experience}) \cdot \beta_1 + (\text{female}) \cdot \beta_2 \\ &+ (\text{education}) \cdot \beta_3 + (\text{female}) \cdot (\text{education}) \cdot \beta_4 + e \end{aligned}$$

Here we assume that the returns to experience is the same for men and women. The covariate (experience) is work experience in years, (education) years of schooling. The covariate (female) is a *gender dummy*, equal to 1 for women and 0 for men. The coefficient β_3 measures the returns to schooling for men, and $\beta_3 + \beta_4$ measures the returns to schooling for women. The null hypothesis is thus that $\beta_4 = 0$.

Another example: you want to see if American cars, Japanese cars and European cars differ in mileage per litre of fuel. You may then specify the model as

$$\begin{aligned} \text{mileage} = \beta_0 &+ \text{engine\_power} \cdot \beta_1 + \text{weight} \cdot \beta_2 \\ &+ (\text{American})\beta_3 + (\text{Japanese})\beta_4 + e. \end{aligned}$$

where (American) and (Japanese) are dummies – equal to one for American cars and zero otherwise, and equal to one for Japanese cars and zero otherwise, respectively. The benchmark is then European cars; β_3 measures the excess mileage of American cars compared to European, and β_4 the excess mileage of Japanese cars compared to European, and the null hypothesis is $\beta_3 = \beta_4 = 0$.

Choice of covariates

In the first case above, maybe we suspect that the returns to schooling is decreasing, i.e., if we already have many years of schooling, one more year have less impact on wage compared to if we are less educated. We can then add the covariate "education squared", for example, and expect the coefficient for this covariate to be negative. Perhaps also a covariate "experience squared" is warranted.

Assume you want to estimate the expected life time of a person, depending on his smoking and alcohol habits. In addition to the natural covariates "smoking" (number of cigarettes per day, for example) and "alcohol consumption", you might want to include the *interaction effect* (smoking)·(alcohol consumption).

Assume you want to run a regression of GNP on some explanatory variables for various countries. Since countries vary a lot in size, it seems reasonable to assume that this causes heavy hetero-skedasticity. Use GNP per capita instead.

Sometimes we don't know which of several covariates should enter the equation. A common test for this is the AIC (Akaike Information Criterion) test. One chooses the model that minimises

$$\mathrm{AIC} = n \ln(|\hat{e}|^2) + 2k$$

where k is the number of coefficients and n the number of observations. In the case of logit, one minimises

$$\mathrm{AIC} = -2\ln(L) + 2k$$

where $\ln(L)$ is the log-likelihood function, i.e.,

$$\ln(L) = \sum_{i=1}^{n} \ln[(2y_i - 1)p(x_i\hat{\beta}) + 1 - y_i].$$

AIC does not provide a test of a model in the sense of testing a null hypothesis. The "true model" – the process that generated the data – is unknown to us, and we wish to select, from among the candidate models, the *approximate* model that minimises the information loss relative to the "true model". This is the one that minimises AIC. We cannot choose with certainty, but we can minimise the *estimated* information loss.

Transformation of variables

In the first example, we used ln(wage) as dependent variable. It is often warranted to use log of the dependent variable if it is positive by nature. For instance, in this case the impact of, say, one more year of experience is that the wage is *multiplied* by $\exp(\beta_1)$, i.e., it increases the wage by a certain *percentage,* rather than a fixed pecuniary amount. Note, however, that if we specify a model

$$\ln(y) = x\beta + e,$$

then the expected value of y is not $\exp(x\beta)$. Indeed, if e is normally distributed with variance σ^2 then

$$\mathrm{E}(y) = e^{x\beta + 0.5\sigma^2}.$$

The *median* value of y is however $e^{x\beta}$.

In some cases this specification can be a remedy for heteroskedasticity. Assume that you consider a model

$$y = x\beta + e$$

and you suspect that the standard error of e is about proportional to $E(y)$. Then we can write this as

$$y = x\beta(1 + v)$$

where the variance of v is essentially independent of x.Taking logarithms gives

$$\ln(y) = \ln(x\beta) + u$$

where u = ln(1+v) whose variance is independent of x. This is thus a homoskedastic equation, and we might consider replacing ln($x\beta$) by the linear specification $x\beta$ to get

$$\ln(y) = x\beta + u.$$

Exercises

1. You want to know how the number of rooms in an apartment influences its price, and consider the two model specifications:

$$\begin{aligned}\text{price} = \beta_0 &+ (\text{no.\_of\_rooms})\beta_1 \\ &+ (\text{floor\_area})\beta_2 + \ldots + \text{resid}\end{aligned}$$

and

$$\text{price} = \beta_0 + (\text{no.\_of\_rooms})\beta_1 + \ldots + \text{resid}$$

(i.e., no floor area in the equation)
Discuss the interpretation of β_1 in the two models.

2. You want to investigate how much wage increase a person may expect if he takes a university exam, compared to if he does not. You run a regression on some data:

$$\begin{aligned}\log(\text{wage}) = \beta_0 &+ (\text{univ\_exam})\beta_1 \\ &+ (\text{work\_experience})\beta_2 + (\text{female})\beta_3 \\ &+ (\text{immigrant})\beta_4 + \text{resid}\end{aligned}$$

Here all variables except (work_experience) are dummies. Do you see any problems with this? If we replace (work_experience) with (age), does that in any way change the interpretation of β_1?

3. A colleague is planning to investigate if a higher tax on beer will serve to decrease traffic fatalities and proposes to use cross-sectional data on (US) states to regress traffic fatalities on *beer tax*, *total miles driven*, *percent of the population that is aged 18–25*, *beer consumption* and *frequency of police road checks*. He will then take the coefficient for *beer tax* as a measure of the impact of tax on beer on traffic fatalities.

   What advice would you give him?

4. You run a regression

$$y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + e$$

   and want to test if $\beta_3 = \beta_4 = 1$. We assume that your programme only can run "vanilla" regressions. How do you perform that test?

VI. Analysis of Observational Data
(econometrics)

Econometrics is concerned with applying statistical methods to problems where available data are *observational* rather than *experimental;* i.e., they don't come from controlled and planned experiments. This is commonly the case in economics, biology, medicine and social sciences. Astronomy is an example of a natural science where data typically are observational. Despite the name, econometrics is thus not confined to the analysis of data in economics.

The first task is to choose a good statistical *model* for the problem under study, and then to perform the statistical analysis of this model. Hence, both *statistical modelling* and *statistical analysis* are important components of applied econometrics.

There are numerous textbooks on econometrics, many very comprehensive. However, they typically focus on the technicalities of estimation, much less on the important issue of modelling. I have tried to put reasonable emphasis on both aspects.

A book that stands out is Peter Kennedy's *A Guide to Econometrics.* It presupposes that you have basic knowledge about the technicalities of econometrics, but explains what is going on, and what can go wrong, and what to do when things look weird, and so on. *This is a book that anyone who is about to do any serious analysis of observational data should read at least relevant chapters from. I most strongly recommend this outstanding and unique book!*

A special feature of observational data that has to be addressed is the possible presence of *endogeneity*.

Endogeneity

The term "endogeneity" is used whenever the assumptions $E(e_i) = 0$ are violated because the expected value of e_i depends on the value of at least one of the covariates; we say that the residual is *correlated with* that covariate. The phenomenon occurs when the regression

equation is given a structural interpretation, not when it is used for prediction. Endogeneity may appear in many different forms, but mathematically the problem is the same. The estimation procedure of OLS requires that the residual is uncorrelated with the covariates, and when this is not the case, *OLS will not produce consistent estimates.* A *positive* correlation will *over estimate* the coefficient, a *negative* correlation will *under estimate* the coefficient (if only one.)

I will go through some common situations when endogeneity arises.

Sample Selection Bias

Bias in data occurs when the probability of being selected to the sample depends on *some other* criterion than the values of the co-variates. For example, if we want to assess if female statistics students perform better or worse than male students, then one of the covariates is a dummy for "female", and hence it is ok to choose an equal number of female students as male students, even if they are not equal in number in classes. It is even commendable to make such a choice. But we must take care so that we don't select especially talented female students, but mediocre male students, for instance, which seems quite obvious.

A common situation is labelled *self selection bias.* Assume that I want to assess if my teaching in class helps students to perform better at the exam. I can then compare the performance of those students that do attend most of my lectures with the performance of those who prefer to study in private. Obviously there is a self selection mechanism at work here. Maybe it is the most talented students who *choose* to study in private. Then the covariate "attendance in class" is correlated with the unobserved characteristic "talent" which is part of the residual.

Simultaneity

"Simultaneity" is present when the alleged "dependent" variable in fact also influences one or more of the covariates, i.e., the cause and effect goes in more than one direction.

Say that we want to assess whether more policemen reduces the

crime rate. So we run a regression of crime rate on some socio-economic variables plus the number of policemen per 100'000 inhabitants. But if the residual is large for some reason, it means that the crime rate is unusually high, and a high crime rate causes authorities to increase the police force. Hence the residual is positively correlated with the covariate "number of policemen".

A more subtle situation is the often given example with "demand" and "supply" of a commodity. Assume I want to estimate the "demand curve" for coffee. The "dependent" variable is then "sold quantity of coffee", and one of the covariates is "price". However, we know from economic theory that if demand goes up, then the price will also go up. Hence, assume that demand goes up at some point in time (some observation,) for instance because of an advertising campaign. This is captured as an residual which is very large (the residual contains "advertising campaign".) Then the price will also go up. Thus, we have a positive correlation between the residual and the covariate "price". (We are in fact measuring a mixture of the demand curve and the supply curve.)

Missing Relevant Covariates

Sometimes the component of the residual that makes it correlate with some covariate can be identified, and in this case we talk about a *missing relevant covariate*. For example, low fuel consumption is a positive feature of a car, so, *ceteris paribus*, low fuel consumption ought to make the car more expensive. However, a regression of price on fuel consumption would probably result in a positive coefficient (higher fuel consumption $\Rightarrow$ higher price). However, a high engine power is also a positive feature which increases both fuel consumption and price. The residual will contain "engine power" which is correlated with the covariate "fuel consumption". In this case the remedy is simple: include the missing covariate "engine power" in the regression.

Measurement Errors

Measurement errors, *even unbiased*, in the covariates cause "endogeneity". Note that (unbiased) measurement errors in the

dependent variable just add a component in the residual, and do not cause "endogeneity". To see why measurement errors in covariates cause endogeneity, consider the "true" model

$$y = x\beta + e$$

and let $\tilde{x} = x + \varepsilon$ be the measured values of x, ε being the measurement error. Then the model we regress is

$$y = \tilde{x}\beta + \tilde{e}\text{, where}$$

$$\tilde{x} = x + \varepsilon \quad \text{and} \quad \tilde{e} = e - \varepsilon\beta$$

Obviously, $\tilde{x}$ and $\tilde{e}$ are correlated unless $\beta = 0$.

A remedy for Endogeneity: Instrumental Variables (2SLS)

We write the model with matrix notation:

$$Y = X\beta + e.$$

When there is endogeneity, i.e., at least one covariate is "endogeneous", meaning that it is correlated with the residual, the by far most common way to handle the situation is by employing *instrumental variables* (or just *instruments*). We must find new variables that are well correlated with the endogeneous ones, but uncorrelated with the residual. The original exogeneous variables *plus* the instruments are called the "*exogeneous variables*", and the original exogeneous variables are the *included exogeneous variables*. The exogeneous variables must be at least as many as the original covariates. In other words, we must find at least one instrumental variable for every endogeneous one.

We denote the matrix of exogeneous variables *Z*. The number of columns in *Z* is hence at least as many as in the matrix *X* of original covariates. If Z has equally many columns as X, then the natural normal equations would be

$$Z'\hat{e} = 0,$$

which is also what they are. However, if Z has more columns than

X, it means that there are more equations in the system above than there are coefficients to be determined, i.e., we have an over determined system. In order to reduce the number of equations such that we get a precisely determined system, we project X onto Z:

$$\hat{X} = Z(Z^t Z)^{-1} Z^t X$$

The normal equations are now

$$\hat{X}^t \hat{e} = 0.$$

(If Z has equally many columns as X, then this is equivalent to the previous system.) The point estimate of β is thus

$$\hat{\beta} = (\hat{X}^t X)^{-1} \hat{X}^t Y,$$

and the heteroskedasticity robust covariance matrix estimator is

$$\mathrm{Cov}(\hat{\beta}) = \frac{n}{n-k-1} (\hat{X}^t X)^{-1} \hat{X}^t D(\hat{e}_i^2) \hat{X} (\hat{X}^t X)^{-1}.$$

The estimated $\hat{\beta}$ is the "Two Stage Least Squares" (2SLS) estimate of β. It is *not* unbiased, but it is consistent.

The approximate F-test and confidence intervals are computed as before.

Weak Instruments

The 2SLS estimator is *asymptotically* unbiased, but biased in finite samples. The OLS estimator is also biased, even asymptotically, but if the 2SLS estimator is not considerably less biased, the OLS estimator is preferred, since its standard errors are smaller. This situation is called *weak instruments*. If we have only one endogeneous variable to worry about, there is a simple test for weak instruments:

Run a regression of the endogeneous variable on the included exogeneous variables of the model, plus the instruments:

$$x = x_1 \gamma_1 + z \gamma_2 + \varepsilon$$

where x is the endogeneous variable, x_1 are the included exogeneous variables of the model and z are the instruments replacing x. Now compute the F-statistic for the hypothesis $\gamma_2 = 0$. This F-statistic should be *large*, at least 10 (recommended by Staiger and Stock;) a smaller value signals a weak instruments problem.

Note that it may be better to have *few* rather than many instruments to avoid weak instruments – contrary to intuition!

If we have more than one endogeneous variable in the model, an appropriate test is more complicated, and we leave that issue here.

Exercises

1. You run a regression

$$y = \beta_0 + (\text{R\&D})\beta_1 + \text{resid}$$

where y is log(GDP/capita) and (R&D) is the expenditure per capita spent on Research and Development. You run the regression on a cross section of countries and get a value for β_1. You want to interpret this value in this way; "If we increase the expenditure on R&D with Δx per capita, then the GDP/capita will will go up by a factor of $\exp(\Delta x\, \beta_1)$." Discuss this interpretation!

2. Assume that you have the following demand – supply system for (retail) coffee:

$$Q^d = \alpha_0 + (\text{retail\_price})\alpha_1 + e_1$$

$$Q^s = \beta_0 + (\text{retail\_price})\beta_1 + (\text{wholesale\_price\_of\_coffee\_beans})\beta_2 + e_2$$

   a) Show that (retail_price) is endogeneous (in both equations).

   b) Show that (wholesale_price_of coffee_beans) is a possible instrumental variable for (retail_price) in the demand equation.

3. A labour economist wants to assess if participation in a labour market programme increases the hazard rate to employment for an unemployed person. He draws a random sample of individuals that were unemployed at a date two years ago and find that they since then all have got a job. He runs the OLS regression:

$$\begin{aligned}\ln(\text{duration}) = {} & \beta_0 + (\text{programme})\beta_1 + (\text{age})\beta_2 \\ & + (\text{experience})\beta_3 + (\text{education})\beta_4 \\ & + (\text{benefit})\beta_5 + (\text{income\_from\_spouse})\beta_6 \\ & + (\text{female})\beta_7 + \text{residual}\end{aligned}$$

   Here "experience" is work experience in years, "education" education in years, "benefit" is a dummy for "eligible for unemployment benefit", "income_from_spouse" and "female" should be self explanatory. The coefficient of interest is of course β_1, since "programme" is a dummy equal to 1 for those who have volunteered to participate in a labour market programme, and equal to zero for those who have chosen not to participate. The economist then takes the estimated value as a measure of the impact of labour market programme on expected unemployment duration.

   Discuss possible problems with this approach.

4. (From Peter Kennedy's book *A Guide to Econometrics*) A Korean friend has regressed the Korean won / US$ exchange rate on its lagged value, the Korean trade balance, the difference between the Korean and US inflation rates and the difference between the Korean and US real interest rates. To his surprise, the coefficient on the trade balance comes out negative, although all other coefficients get the expected sign.

   Can you explain to him what is going on? What remedy do you suggest to him? (You might need some basic insight in macro economics for this exercise.)

5. You want to see if males and females differ after three years of study at KTH:s programme on computer science. The issue is if they differ in the number of credit points they have managed to collect, on average.

   For the purpose, you collect data on 50 male students and 50 female students, and run the regression

$$(\text{credit\_points}) = \beta_0 + (\text{female})\beta_1 + e$$

   (female) is a dummy for female.

   a) The fraction of females at KTH:s computer science programme is only 15%. You have chosen 50 males and 50 females, hence not a random sample of students. Explain why this *does not* cause a "selection bias" in this case.

   b) Prove that if you had chosen a random sample of students, then the standard error of the estimated β_1 would presumably be greater by a factor of about 1.4. (It is hence a good idea to *not* sample the students randomly.)

VII. Analysis of Experimental Data
(Experimental Design and ANOVA)

Now we look 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 factor 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 hours 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. Smaller residuals give smaller standard errors of the coefficients. Hence, we would like to keep this variation as small as possible in order to make a better inference about the size 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" (see exercise 2 further down) and "incomplete block designs". I refer to the literature for these.

One-way ANOVA

When an experiment has been set up, and the data from the appropriate experiments have been collected, we want to use these data to make inferences about the population means associated with the various treatments. The method used to compare the treatment means is traditionally known as *analysis of variance*, or ANOVA. The formulas – one set for each experimental design – were developed in the early 1900, well before the invention of computers. These formulas can be seen as special cases of linear regressions, where the response variable is regressed on dummies for the various populations and treatments. This is the approach we will use here.

We begin with an example of a *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. We take several measurements at each location:

A: 42, 37, 41, 39, 43, 41

B: 37, 40, 39, 38, 41, 39

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 ("homoskedasticity"). Now we want to test the hypothesis H_0: $\mu_A = \mu_B = \mu_C$ against the opposite H_1: "at least two population means differ". We first run a regression of the measurements on dummies for location *B* and *C*, i.e., we take location *A* as reference point:

$$y = \beta_0 + (B)\beta_1 + (C)\beta_2 + e$$

The estimated coefficients are thus

$$\hat{\beta}_0 = \hat{\mu}_A, \quad \hat{\beta}_1 = \hat{\mu}_B - \hat{\mu}_A, \quad \hat{\beta}_2 = \hat{\mu}_C - \hat{\mu}_A.$$

The estimated differences are now computed as

$$\hat{\mu}_B - \hat{\mu}_A = -1.5, \quad \hat{\mu}_C - \hat{\mu}_A = -9.0, \quad \hat{\mu}_C - \hat{\mu}_B = -7.5.$$

The last difference is of course obtained as $\hat{\beta}_2 - \hat{\beta}_1$. The null hypothesis that all population means are equal is equivalent to $\beta_1 = \beta_2 = 0$, whose *F*-statistic is 32.056, computed with our heteroskedasticity consistent formula. However, the tradition in ANOVA analysis is to assume homoskedasticity, and the computed *F*-statistic is in this case

$$F = \frac{R^2}{1-R^2}\frac{n-k-1}{r}$$

where the notation is as before: n = total number of observations

(= 18), k = number of covariates (= 2) and r = number of coefficients tested for zero (= 2). In this case $F = 36.71$. This F-statistic has an exact $F(r, n-k-1)$ distribution under the hypothesis "all observations are independent draws from the same normal distribution".

We summarise the result in an ANOVA table (residual $df = n-k-1$):

source	df	F	p-value	η^2
location	2	36.71	0.0000	0.83
residual	15			

In the current case, we of course reject the hypothesis H_0 in favour for H_1, i.e., the population means are not all equal. If we had not rejected H_0, the analysis would have ended here.

Multiple comparisons

Since we accept the hypothesis that *not* all μ_k:s are equal, it is natural to investigate which pairs that differ. One way to do this is to compute confidence intervals for each difference. The individual standard errors for these differences are computed under the assumption of homoskedasticity. This is done as follows: First we compute the standard error of the regression, i.e., an estimate of SD(e):

$$s = \sqrt{\frac{|\hat{e}|^2}{n-k-1}}$$

(This number is always reported by the computer software in the output of the regression.) The standard errors are now

$$\mathrm{SE}(\hat{\mu}_B - \hat{\mu}_A) = s\sqrt{1/n_B + 1/n_A}\ ,$$

etc, where n_A is the number of measurements at location A, etc.

According to Bonferroni, if we want to use a total error rate (the "error" being that at least one interval does not contain the true

value) of, say, 5%, then we should use a confidence level of $1 - 0.05/3$ for each interval.

In the current case, $s^2 = 3.8$, so each standard error is

$$\sqrt{3.8\,(6^{-1} + 6^{-1})} \approx 1.1255$$

so the confidence intervals are

$$\text{corresponding point estimate} \pm 2.6937 \cdot 1.1255$$

(here 2.6937 is the square root of the 0.05/3 quantile of the $F(1, 15)$ distribution.) We summarise in a table:

Confidence intervals, experimental confidence level 95% (Bonferroni)

$$\begin{aligned} \mu_B - \mu_A &= -1.5 \pm 3.03 \\ \mu_C - \mu_A &= -9.0 \pm 3.03 \\ \mu_C - \mu_B &= -7.5 \pm 3.03 \end{aligned}$$

We conclude that $\mu_C < \mu_A$ and $\mu_C < \mu_B$.

There are other methods for multiple comparisons, as Sheffé's test and Tukey's method. We refer to the literature for these.

Two-way ANOVA

Assume that we want to asses how the the humidity of paper, produced by a paper machine, varies with the level of speed of the machine and the use of ingredients. Here is an experiment: Two measurements were made for each of two levels of speed, and three different mixture of ingredients. The results are displayed here:

	ingredients 1	ingredients 2	ingredients 3
speed 1	7.2 | 7.2	7.8 | 7.2	8.4 | 7.8
speed 2	6.4 | 6.8	7.2 | 7.2	7.8 | 7.6

We first consider an additive model: let y be the humidity when speed i is applied to ingredients j and assume that

$$y = \alpha + x_2\beta + z_2\gamma_2 + z_3\gamma_3 + e.$$

Here x_2 is a dummy for speed 2, , z_2 and z_3 dummies for ingredients 2 and 3, and e is a random residual which we assume is $N(0,\sigma)$. The benchmark case is thus speed 1 with ingredients 1. We run this regression, and save the output s^2, the residual standard error (computed as above.) We want to test if

i) speed has an impact on humidity, and

ii) ingredients have an impact on humidity.

We need the F-statistics for

i) $\beta = 0$, and

ii) $\gamma_2 = \gamma_3 = 0$.

Since we assume homoskedasticity, we will compute these F-statistics accordingly. There are more than one way to do this, but the simplest way in the current context is

$$F = \frac{\eta^2}{1-\eta^2}\frac{n-k-1}{r},$$

where, as before, n = total number of observations (= 12), k = number of covariates (= 3) and r = number of coefficients tested for zero (=1 and 2). The F-statistic follows an $F(r, n-k-1)$ distribution under the null.

We summarise in an ANOVA table:

source	*df*	F	p-value	η^2
speed	1	8.89	0.0175	0.53
ingredients	2	15.84	0.0017	0.80
residual	8			

The *p*-values are now individual *p*-values. If we employ Bonferroni's method, we should use twice these *p*-value as the experimental *p*-value, i.e. the risk of making rejecting at least one true hypothesis. For instance, if we accept an experimental error rate of 2%, we can not reject the hypothesis that speed has no impact on humidity, since 2·0.0175 exceeds 2%.

What does Akaike say? Does Akaike prefer the smaller model with only ingredients and no speed factor? As is readily shown, Akaike prefers the smaller model if (and only if)

$$\eta^2 < 1 - e^{-2r/n}.$$

In this case $1 - e^{-2r/n} = 0.2835$ for $r = 2$ and $= 0.1535$ for $r = 1$. Hence, Akaike prefers the full model.

In order to decide whether we should reduce a model or not, employing the AIC criterion might be of help, but ultimately it is a subjective decision. In any case, it is usually advisable to reduce the model if 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 smaller. If we have an equal number of replications (two, in the current example), the estimated coefficients will be exactly the same, but the errors of these estimates should change slightly. (If $F < 1$, they will go down, since it means that the regression standard error has decreased, so it was presumably a good idea to reduce the model.)

The coefficient γ_2, for instance, is of course the difference between speed 2 and speed 1, and a confidence interval is computed as before (ii.2), but with the homoskedastic covariance matrix (iv.1) employed.

Interactions

The previous model is an *additive* one; the difference between the mean humidity at speed 1 and speed 2 is constant across ingredients. A model that gives each combination of speed and ingredients a unique mean value is one where *interaction effects* are incorporated:

$$y = \alpha + x_2\beta + z_2\gamma_2 + z_3\gamma_3 + x_2z_2\delta_{22} + x_2z_3\delta_{23} + e.$$

Here the product x_2z_2 is a dummy for "speed 2 with ingredients 2" etc., the benchmark is still "speed 1 with ingredients 1". The ANOVA table now looks like this:

source	*df*	*F*	*p*-value	η^2
main	3	10.17	0.0091	0.84
interactions	2	0.304	0.7484	0.092
residual	6			

Here only the row with "interactions" is of any interest. The "main" row pertains to the coefficients β, γ_2 and γ_3, i.e., the "main" effects. However, the *F*-value, *p*-value and η^2 depend on which benchmark we use (speed 1 with ingredients 1 in this case.) Let us look at an example:

Assume that we have two factors, each at two levels. The first has levels *Low* and *High*, the second *Cold* and *Hot*. Assume that the true model is

$$y = 4 + (High \cdot Hot) \cdot 4 + e$$

i.e., the response is 4 if we employ *Low* and *Cold*, and if we increase *Low* to *High*, nothing happens, the response is still 4, and if we keep *Low* and increase temperature to *High*, nothing happens either; the response is still 4. But if we do *both* – we change from *Low* to *High* and from *Cold* to *Hot*, the the response goes up to 8. In this case, there are no "main" effects, only an interaction effect (both levels have to be changed.)

However, in traditional ANOVA modelling, the equation would look like this:

$$y = 5 - Low + High - Cold + Hot + (Low \cdot Cold) - (Low \cdot Hot) - (High \cdot Cold) + (High \cdot Hot) + e$$

You can check that the two equations are equivalent: $y = 4$ (on average) for all configurations except for *High* + *Hot* which yields $y = 8$. But in this case there seems to be "main effects".

So in order to not give some false impression, I suggest the following table

source	*df*	*F*	*p*-value	η^2
main	3	–	–	–
interactions	2	0.304	0.7484	0.092
residual	6			

The only thing we use the table for is to decide if we should exclude the interaction effects from the model. If we *don't*, then we keep also all main effects, whereas if we *do* exclude the interactions, then we *re-estimate the model as an additive one.*

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 18 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 18 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.

VIII. ANOVA on Ranked Data

It happens that our dependent data are far from normally distributed, or that they are just ordered, but not quantitative, like "very bad" < "rather bad" < "rather good" < "very good". In this case it is common practise to rank the data, and then regress them on dummies for categories.

Let us do this on the Mississippi River data we used earlier. The lowest value, 28, is given the rank 1, the second lowest, 30, is given the rank 2 etc,. and the largest value, 43, is given the value 18. Ties are given their mean ranks, such that 32 has rank 3.5, 37 the rank 7.5, and 39 the rank 11. We regress these ranks on dummies (for instance B, C and an intercept, using A as reference.) The resulting $R^2 = 0.72593$. The test variable is now $Q = R^2(n-1) = 0.72593 \cdot 17 = 12.34$, where n is the number of observations. If all true slope coefficients are zero (i.e., if the true mean ranks are the same in each category,) then Q is approximately an observation of a $\chi^2(r-1)$ variable, where r is the total number of categories, i.e., $r = 3$ in this case. Since $\Pr(\chi^2(2) > 12.34) = 0.0021$, it is reasonable to reject that hypothesis.

This test is called the *Kruskal-Wallis test*, or the *H-test*. In the particular case $r = 2$, it is called *Wilcoxon's rank-sum test*, the *U-test* or the *Mann-Whitney test*.

The reason that we have a χ^2 statistic here, rather than the earlier F-statistic, is that under the null hypothesis, we know the exact variance of the residual. One can alternatively employ the F-statistic with heteroskedasticity correction, and some statisticians claim that this is more efficient as regards type two errors, but since Wilcoxon's and Kruskal-Wallis' tests are so established, I chose to put them forth here.

The test can sometimes be used for data in a contingency table, as an alternative to Pearson's chi-square test. Here is an example:

In Connellan, J. et.al., *Sex differences in human neonatal social perception*, Infant Behavior & Development 23 (2000), the researchers write (Abstract)

> "*Sexual dimorphism in sociability has been documented in humans. The present study aimed to ascertain whether the sexual dimorphism is a result of biological or socio-cultural differences between the two sexes. 102 human neonates, who by definition have not yet been influenced by social and cultural factors, were tested to see if there was a difference in looking time at a face (social object) and a mobile (physical-mechanical object). Results showed that the male infants showed a stronger interest in the physical-mechanical mobile while the female infants showed a stronger interest in the face. The results of this research clearly demonstrate that sex differences are in part biological in origin.*"

Let us look at the data they use in this report. I refer to the full report for details. In short, the 102 neonates were presented to look at a human face and a mechanical mobile, and their interest in each of these were measured. Here is a presentation of the data:

	face preference	no preference	mobile preference
males (44)	11	14	19
females (58)	21	27	10

A χ^2-test (contingency table) gives the p-value $p = 0.016$ for the hypothesis of homogeneity, so this hypothesis is rejected (at 2% risk level.) Hence, there is a gender difference related to these preferences. This is the presentation given in the article.

However, the test doesn't tell us in what way males and females differ. Let us rank "face preference" < "no preference" < "mobile preference". Then "face preference" gets the rank 16.5, "no preference" the rank 53 and "mobile preference" the rank 88. We regress these 102 ranks on an intercept (males) and a dummy for

females. The coefficient for the female dummy comes out negative (-13.17), $R^2 = 0.05569$. Hence, $Q = 0.05569 \cdot 101 = 5.625$, so the *p*-value for this coefficient be zero is $\Pr(\chi^2(1) > 5.625) = 0.0177$. Hence, we conclude that males have, on average, more interest in the mobile compared to the females.

The regression coefficient β for the female dummy has a clear interpretation: If we draw one male and one female at random, then

$$\Pr(\text{female} > \text{male}) - \Pr(\text{female} < \text{male}) = 2\beta/(44+58) = -0.258.$$

Here 44 and 58 are the number of males and females, and the inequality female > male means that the female is more interested in the mobile than the male is, i.e., the female gets the higher rank.

Why do we assume homoskedasticity in ANOVA?

Heteroskedasticity is a major practical issue encountered in ANOVA analyses. We have used the *heteroskedasticity robust covariance matrix* suggested by Halbert White in this text, but unfortunately it doesn't work well when there are few residual degrees of freedom, as is always the case in ANOVA analyses. In fact, in the two-way ANOVA example with interactions above, the heteroskedastic covariance matrix (ii.1) is singular, and the estimated variance for the intercept is zero.

Several fixes have been suggested for the *F*-test, but as of yet there seems not to have emerged any consensus about any.

As for differences in population means, a natural method – in my opinion – is to employ Welch's unequal variance *t*-test, which is implemented in all statistical software, and also in (most) statistical pocket calculators. One then compares each pair of groups *in isolation*, where data from the other groups do not enter. Note that the traditional method uses data from *all* groups when s^2 is computed, so for instance in our Mississippi example, when the confidence interval for $\mu_B - \mu_A$ computed, also data from location *C* are involved. One can imagine that if the variance at location *C* is different from that in *A* and *B*, this introduces an error in the

estimate. It seems to me that pairwise comparisons with Welch's method, together with Bonferroni, often is a reasonable substitute for one-way ANOVA.

Five Commandments of Applied Statistics

In his book "A Guide to Econometrics" Peter Kennedy gives ten commandments for applied econometrics. I will cite five of them here:

Thou shalt not worship complexity
: Complex methods are often less robust to specification errors. And all models are to some extent misspecified.

Thou shalt not confuse significance with substance
: A large *p*-value can be due to a large standard deviation of the estimate; the impact can still be substantial. Similarly, a small *p*-value can be the result of a very precise estimate (e.g., many data); the impact can still be negligible.

Thou shalt ask the right questions
: Put relevance before elegance. I.e., it is better to make an approximate analysis of the right model rather than an exact analysis of an irrelevant.

Thou shalt know the context
: You must be familiar with the phenomenon being investigated in order to perform a valid statistical analysis.

Thou shalt beware the costs of data mining
: "Don't hunt for statistical significance with a shotgun."

Exercises

1. In an experiment designed to evaluate two types of detergents, *A* and *B*, a laboratory ran three loads of washing at each combination of detergents and water temperatures, and obtained the following whiteness readings:

	detergent *A*			detergent *B*		
low temp.	49	48	50	43	41	40
high temp.	46	50	50	44	38	42

   Do an ANOVA analysis and test for differences between the readings due to detergents, temperature and interaction.

2. (**Latin Squares**) Assume that you want to test three types of glue ("treatments"), *A*, *B*, *C*, under tree different temperatures, T1, T2, T3 and three pressures P1, P2, P3. If you want to do no more than nine experiments, there are at least two possible experimental designs to consider; see display below.

	P1	P2	P3
T1	*A*	*B*	*C*
T2	*B*	*C*	*A*
T3	*C*	*A*	*B*

	P1	P2	P3
T1	*A*	*A*	*C*
T2	*B*	*A*	*B*
T3	*B*	*C*	*C*

   In the left design, we vary two factors between each reading of glue (A with factors P1,T1 and P3,T2 and P2,T3, etc.,) whereas we in the right vary only one factor at a time (A with factors P1,T1 and P2,T1 and P2,T2.) Prove that the left design will give much smaller (true) variances of the estimated coefficients. In fact, the variances of the "treatments" (glue) will be three times as large in the right design, and the "factor" (pressure and temperature) variances will be twice as large.

   The left, which is obviously the most efficient, is called a *Latin Square* design.

3. In a randomised block design groups of homogeneous "experimental units" are employed in order to compare the means of the populations associated with a number of "treatments". The analysis is a two way ANOVA with no interactions. Perform the analysis in the following example:

   Prior to submitting a bid for a construction job, cost engineers prepare a detailed analysis of the estimated labour and materials cost required to complete the job.

   A company that employs three job cost engineers wanted to compare the mean level of the engineer's estimates. Each engineer estimated the same four randomly selected jobs (jargon: the jobs are the "blocks", the engineers the "treatments".) Here are the data:

	job 1	job 2	job 3	job 4
engineer 1	4.6	6.2	5.0	6.6
engineer 2	4.9	6.3	5.4	6.8
engineer 3	4.4	5.9	5.4	6.3

   Perform an ANOVA analysis and test to determine if there is sufficient evidence for differences among treatment means. Use error risk 0.05.

   Compute also the effect size η^2 for the blocking on jobs to assess how successful it was in reducing the job-to-job variation in the estimates.

4. 205 males and 195 females were asked to what extent they were in favour of a certain community arrangement. *Is this arrangement a good or bad idea?* The result was as follows:

	very bad	rather bad	rahter good	very good
males	48	91	40	26
females	59	88	32	16

We want to assess if there is a difference between genders as to how favourable they are to the arrangement (i.e., are males more favourable to the arrangement compared to females, or vice versa?) State a precise hypothesis, perform the test, and decide what you can accept or not, at an error rate of 5%.

IX. Appendix.

Some proofs and derivations

1. Tests and confidence intervals

If we have a reasonable amount of data, then the estimated coefficients $\hat{\beta}$ will be approximately normally distributed (Central Limit Theorem) with mean value β. Hence

$$\frac{(\hat{\beta}_j - \beta_j)^2}{\sigma_j^2}$$

will have an approximate $\chi^2(1)$ distribution. Here σ_j^2 is the variance of $\hat{\beta}_j$. However, this variance is estimated from data, and we compensate for this by employing the $F(1, n - k - 1)$ distribution instead. (This is the exact distribution if the model is homoskedastic and estimated accordingly, and the residuals are normal.)

By the same token, if $\hat{\beta}_j$ is an r-subvector of $\hat{\beta}$, then

$$\hat{\beta}_j^t V_j^{-1} \hat{\beta}_j$$

– where V_j is the covariance matrix – has an approximate $\chi^2(r)$ distribution. Again, since V_j is estimated, we compensate for this by employing the $F(r, n - k - 1)$ distribution (with a factor $1/r$) instead. (Also in this case, this is the exact distribution if the model is homoskedastic, the residuals are normal and $\hat{V}_j$ is computed from the covariance matrix (iv.1).)

2. R^2 and Effect Size

The definition of η^2 is

$$\eta^2 = \frac{|\hat{e}_*|^2 - |\hat{e}|^2}{|\hat{e}_*|^2}$$

and the proof for the alternative formula (ii.3)

$$\eta^2 = \frac{q}{|\hat{e}|^2 + q}$$

is somewhat tricky. But here goes! We use the notation from the section "R^2 and Effect Size". First we regress the x_2 variables on x_1:

$$x_2 = x_1\hat{\Gamma} + \hat{u} \qquad \text{(A.2.1)}$$

(each component of x_2 is thus regressed on x_1.)

Here $X_1^t\hat{U} = 0$ (normal equations.) Hence, it follows that

$$X^tX = \begin{pmatrix} X_1^t \\ \hat{\Gamma}^t X_1^t + \hat{U}^t \end{pmatrix}\begin{pmatrix} X_1 & X_1\hat{\Gamma} + \hat{U} \end{pmatrix}$$

$$= \begin{pmatrix} X_1^tX_1 & X_1^tX_1\hat{\Gamma} \\ \hat{\Gamma}^tX_1^tX_1 & \hat{\Gamma}^tX_1^tX_1\hat{\Gamma} + \hat{U}^t\hat{U} \end{pmatrix}$$

and with some calculations one can see that it follows that

$$(X^tX)^{-1} = \begin{pmatrix} A & B \\ B^t & (\hat{U}^t\hat{U})^{-1} \end{pmatrix}$$

for some matrices *A* and *B*. Hence,

$$q = \hat{\beta}_2^t\hat{U}^t\hat{U}\hat{\beta}_2 = \left|\hat{U}\hat{\beta}_2\right|^2.$$

Now we go back to the regression equation (A.2.1). We know that $X_1^t\hat{U} = 0,$ but also that $X_1^t\hat{e} = 0$ (normal equations again,) hence $X_1^t(\hat{U}\hat{\beta}_2 + \hat{e}) = 0$. Substituting (A.2.1) into the original equation yields

$$y = x_1(\hat{\beta}_1 + \hat{\Gamma}\hat{\beta}_2) + \hat{u}\hat{\beta}_2 + \hat{e}$$

and since $X_1^t(\hat{U}\hat{\beta}_2 + \hat{e}) = 0$ the normal equations are satisfied, so this is the regression result when y is regressed on x_1, and $\hat{U}\hat{\beta}_2 + \hat{e}$ is the residual $\hat{e}_*$. Now $X_1^t\hat{e} = 0$ and $X_2^t\hat{e} = 0$, and since $\hat{U}$ is a linear

combination of X_1 and X_2, it follows that $\hat{U}^t\hat{e} = 0$. By Pythagoras' theorem

$$|\hat{e}_*|^2 = |\hat{U}\hat{\beta}_2|^2 + |\hat{\mathrm{e}}|^2, \text{ i.e., } \quad |\hat{e}_*|^2 - |\hat{\mathrm{e}}|^2 = |\hat{U}\hat{\beta}_2|^2 = q.$$

It now follows that

$$\frac{q}{|\hat{\mathrm{e}}|^2 + q} = \frac{|\hat{e}_*|^2 - |\hat{\mathrm{e}}|^2}{|\hat{e}_*|^2} = \eta^2,$$

Q.E.D.

3. Estimation of σ^2

Note that

$$\begin{aligned}\hat{e} &= Y - X\hat{\beta} = Y - (X^tX)^{-1}X^tY \\ &= (I - X(X^tX)^{-1}X^t)(X\beta + e) \\ &= (I - X(X^tX)^{-1}X^t)e\end{aligned}$$

Here

$$\mathrm{Tr}(X(X^tX)^{-1}X^t) = \mathrm{Tr}((X^tX)^{-1}X^tX) = \mathrm{Tr}(I_{k+1}) = k+1.$$

Let $P = (I - X(X^tX)^{-1}X^t)$, then $\mathrm{Tr}(P) = n - k - 1$ and

$$\begin{aligned}\mathrm{E}(|\hat{e}|^2) &= \mathrm{E}(\hat{e}^t\hat{e}) = \mathrm{E}(\mathrm{Tr}(\hat{e}\hat{e}^t)) \\ &= \mathrm{E}[\mathrm{Tr}(Pee^tP)] = \mathrm{Tr}(P(I\sigma^2)P) \\ &= \mathrm{Tr}(P^2)\sigma^2 = \mathrm{Tr}(P)\sigma^2 = (n-k-1)\sigma^2\end{aligned}$$

Hence, an unbiased estimate of σ^2 is

$$s^2 = \frac{1}{n-k-1}|\hat{e}|^2$$

4. Proof that a positive correlation between the residual and a covariate will over estimate that coefficient.

The equation is

$$y = x_1\beta_1 + x_2\beta_2 + e$$

where x_2 (a single covariate) is correlated with e. Regress x_2 on x_1:

$$x_2 = x_1\hat{\gamma} + \hat{u} \qquad \text{(A.4.1)}$$

The covariance $\mathrm{Cov}(x_2, e) = \mathrm{Cov}(\hat{u}, e) = \mathrm{E}[\hat{u}^t e]$. Now

$$y = x_1(\beta_1 + \hat{\gamma}\beta_2) + \hat{u}\beta_2 + e\,,$$

hence (normal equations)

$$\hat{u}^t y = |\hat{u}|^2 \beta_2 + \hat{u}^t e, \qquad \text{i,e.,} \qquad \beta_2 = \frac{\hat{u}^t y}{|\hat{u}|^2} - \frac{\hat{u}^t e}{|\hat{u}|^2}.$$

But by the same token, since $\hat{u}^t \hat{e} = 0$,

$$\hat{\beta}_2 = \frac{\hat{u}^t y}{|\hat{u}|^2}, \quad \text{i.e.,} \quad \beta_2 = \hat{\beta}_2 - \frac{\hat{u}^t e}{|\hat{u}|^2}.$$

It now follows that if $\mathrm{E}[\hat{u}^t e] > 0$, then β_2 will be over estimated, etc.

5. On Multicollinearity

We use the notation from appendix 4. Let R_2^2 be the coefficient of determination for equation (A.4.1), which is also the square of the correlation coefficient between x_2 and the linear combination $x_1\hat{\gamma}$. Hence, if R_2^2 is close to 1, that indicates multicollinearity. The calculations in appendix 2 give, after some rearrangements,

$$\mathrm{SE}(\hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}^2}{(n-1)\hat{V}(x_2)}}\,\frac{1}{\sqrt{1-R_2^2}}$$

where $\hat{V}(x_2)$ is the sample variance of x_2 and $\hat{\sigma}$ the standard error of the residual.

Note that if R_2^2 is close to 1, then $\mathrm{SE}(\hat{\beta}_2)$ is large, and the factor $(1-R_2^2)^{-1}$ is called the *Variance Inflation Factor* (VIF). In some (older) literature, this factor is taken as a measure of multi-collinearity, and one can see the rule of thumb that VIF > 10 signals a problem.

But, as we see from the formula, a large value of $(1-R_2^2)^{-1}$ is compensated by a large value of n. In fact, a large value of $(1-R_2^2)^{-1}$ has exactly the same consequence as a small value of n, i.e., few observations ("micronumerosity".)

6. Confidence interval for η^2

The "true" value of η^2 is defined as follows (notation from "R^2 and Effect Size"): let

$$x_2 = x_1\hat{\Gamma} + \hat{u}, \text{ then}$$

$$\eta^2 = \frac{\left|\hat{u}\beta_2\right|^2}{\left|\hat{u}\beta_2\right|^2 + \left|e\right|^2}$$

with obvious notation. Let F be the (homoskedastic) F-statistic for the null hypothesis $\beta_2 = 0$. Then F is an observation of an $F(r, n-k-1, \lambda)$ distributed random variable, where λ is the *non-centrality* parameter. This parameter is related to η^2 as

$$\eta^2 = \frac{\lambda}{n+\lambda}.$$

Hence we can compute e.g. a 95% confidence interval for η^2 as

$$\frac{\lambda_L}{n+\lambda_L} < \eta^2 < \frac{\lambda_U}{n+\lambda_U}$$

where λ_L and λ_U are computed from

$$\Pr(F(r, n-k-1, \lambda_L) \geq F) = 0.025 \qquad \text{and}$$

$$\Pr(F(r, n-k-1, \lambda_U) \leq F) = 0.025.$$

The effect size η^2 as defined in the text has a tendency to over-estimate the "true" value as defined above. We can see this for example in the ANOVA table on p.41. The value 0.092 for the interaction corresponds to $\lambda = 1.216$. But $\Pr(F(2, 6, 1.216 > 0.304) \approx$ 0.85, so (with some abuse of language) with 85% probability, 0.092 is an over-estimate.

An "unbiased" estimate is $\eta^2 = \dfrac{\lambda}{n+\lambda}$, where λ is chosen such that $\Pr(F(r, n-k-1, \lambda) \leq F) = 0.5$. This estimate is "unbiased" in the sense that it is an over-estimate with probabbility 0.5, and an under-estimate with probabbility 0.5.

INDEX