Experimental Design and Biostatistics (TAMS38)

Lecture 9 – Choice of sample size - Linear regression analysis

Martin Singull

Department of Mathematics
Mathematical Statistics
Linköping University, Sweden

LINKÖPING UNIVERSITY

Content

- Power of the test
- Choice of the sample size
- Repetition – Multiple Linear regression

Choice of the sample size

When planning a sample survey or an experiment one would like to choose sample size so that survey becomes "good".

This is accomplished by

a) requirements on the length of a given confidence interval,
b) requirements on the significance level and power.

In both cases we need often information about unknown parameters such as for example standard deviation.

Example 1 – Choice of the sample size

Case a) We have observations $x_1, \ldots, x_n$ from $N(\mu, \sigma)$, i.e.,

$$X_i = \mu + \varepsilon_i \sim N(\mu, \sigma).$$

We want to get information about μ and about unknown standard deviation σ. In the usual way we can obtain the confidence interval for μ with confidence level $1 - \alpha$:

$$I_\mu = \left(\bar{x} - t \cdot \frac{s}{\sqrt{n}} \, , \, \bar{x} + t \cdot \frac{s}{\sqrt{n}} \right),$$

where t is given by $t(n-1)$-table such that $F(t) = 1 - \alpha/2$.

Suppose we require that interval length shall not exceed a.

The interval length is

$$|I_\mu| = \bar{x} + t \cdot \frac{s}{\sqrt{n}} - \left(\bar{x} - t \cdot \frac{s}{\sqrt{n}} \right) = 2t \cdot \frac{s}{\sqrt{n}}.$$

Hence, we have

$$\frac{ts}{\sqrt{n}} \leq \frac{a}{2} \iff n \geq \left(\frac{2ts}{a} \right)^2.$$

For $\alpha = 0.05$ we have $t \approx 2$ and then the sample size $n \gtrapprox \left(\frac{4s}{a} \right)^2$.

Now, we need some "guess" about s^2.

Choice of the sample size, cont.

Case b). Let us now suppose that we have observations $x_1, \ldots, x_n$ from $N(\mu, \sigma)$ and that we want to test

$$H_0 : \ \mu = \mu_0 \quad \text{versus} \quad H_1 : \ \mu \neq \mu_0$$

with help of the test on **significance level** α.

Furthermore, we want that test gives us verdict "reject H_0" with probability $1 - \beta$ if μ_1 is the true μ-value, i.e., the **power should** $1 - \beta$ for $\mu = \mu_1$.

Hence, we have two conditions

(i) $$\alpha = P(H_0 \text{ is rejected if } H_0 \text{ is true})$$

(ii) $$1 - \beta = P(H_0 \text{ is rejected if } \mu = \mu_1)$$

1. Using the conditions (i) and (ii) and the guessed value of σ one can calculate the necessary sample size, see Example and the Kirkwood table with formulas.
2. For the given sample size n and given significance level α one can simulate the power of test for the given value of σ.
3. Minitab has an available routine "Power and sample size".

Example 2 – Single mean

One wants to investigate PCB levels in fish from a certain lake. We catch n fishes and for each of them measure the PCB level to obtain the observations $x_1, \ldots, x_n$.

Model: $X_i = \mu + \varepsilon_i$ where $\mu =$ "true PCB level of fishes in lake" and ε_i describes the variation, $\varepsilon_i \sim N(0, \sigma)$.

Since before, we had $\sigma = 0.9$ for the fishes and we assume that this value is still correct.

One wants to test

$$H_0 : \ \mu = 10.8 \quad \text{against} \quad H_1 : \ \mu \neq 10.8$$

with help of the test on the significance level $\alpha = 5\%$ and with power $1 - \beta = 90\%$ for $\mu_1 = 11.5$. Hence, we have $\mu_0 = 10.8$.

We have $\hat{\mu} = \bar{x}$; r.v. $\overline{X} \sim N\left(\mu, \frac{0.9}{\sqrt{n}}\right)$.

Test statistics: $z = \dfrac{\bar{x} - 10.8}{0.9/\sqrt{n}}$

The r.v. $Z \sim N(0, 1)$ if H_0 is true.

H_0 is rejected if $z < -v$ or $z > v$. Table gives $v = 1.96$ because $\Phi(v) = 0.975$.

H_0 is rejected if

$$\bar{x} < 10.8 - v \cdot \frac{0.9}{\sqrt{n}} = c_1 \quad \text{or} \quad \bar{x} > 10.8 + v \cdot \frac{0.9}{\sqrt{n}} = c_2$$

If H_0 is true then $\overline{X} \sim N\left(10.8, \frac{0.9}{\sqrt{n}}\right)$. Therefore, H_0 is rejected when $\bar{x}$, value is far out in the tail of the normal distribution.

Power for $\mu = 11.5$ is

$$
\begin{aligned}
1 - \beta &= P(H_0 \text{ is rejected given } \mu = 11.5) \\
&\approx P\left(\overline{X} > 10.8 + v \cdot \frac{0.9}{\sqrt{n}} \text{ given } \mu = 11.5\right) \\
&= P\left(\frac{\overline{X} - 11.5}{0.9/\sqrt{n}} > \frac{10.8 + v \cdot \frac{0.9}{\sqrt{n}} - 11.5}{0.9/\sqrt{n}} \text{ given } \mu = 11.5\right) \\
&= 1 - \Phi\left(-\frac{0.7}{0.9/\sqrt{n}} + v\right) = 1 - \Phi(-u).
\end{aligned}
$$

We see that $\Phi(u) = 1 - \beta = 0.90$, so $u = 1.28$.

Then

$$-u = -\frac{0.7\sqrt{n}}{0.9} + v \iff \sqrt{n} = \frac{(u+v)\cdot 0.9}{0.7} \iff$$
$$n = \frac{(u+v)^2 \cdot 0.9^2}{0.7^2} = 17.35$$

and the condition for the sample size become $n \geq 18$.

Case 1 on "Kirkwoods list":

$$n \geq \frac{(u+v)^2\sigma^2}{(\mu-\mu_0)^2}$$

We had $u = 1.28$, $v = 1.96$, $\sigma = 0.9$, $\mu - \mu_0 = 11.5 - 10.8 = 0.7$.

Remark. Value $u + v$ is sometimes called "power index" (PI).

Example 2, cont. - One-sided test

If the alternative hypothesis is instead in the form $H_1: \mu > 10.8$ then H_0 is rejected if

$$z = \frac{\bar{x} - 10.8}{0.9/\sqrt{n}} > v'$$

Table gives $v' = 1.645$ ($\Phi(v') = 1 - \alpha$).

Power calculations give

$$n' \geq \frac{(u + v')^2 \cdot \sigma^2}{(\mu - \mu_0)^2} = \frac{(u + v')^2 \cdot 0.9^2}{0.7^2} = \frac{(1.28 + 1.645)^2 \cdot 0.9^2}{0.7^2} = 14.14$$

Hence $n' \geq 15$.

Comparison of two means

Observations $x_1, \ldots, x_n$ from $N(\mu_1, \sigma_1)$ and $y_1, \ldots, y_n$ from $N(\mu_2, \sigma_2)$, where σ_1 and σ_2 is known.

We want to test

$$H_0 : \ \mu_1 = \mu_2 \quad \text{against} \quad H_1 : \ \mu_1 \neq \mu_2.$$

Test statistics

$$\frac{\bar{x} - \bar{y} - 0}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}}$$

Corresponding r.v. is $N(0, 1)$ under H_0 and we obtain, by similar calculations as above

Case 4 on "Kirkwoods list":

$$n \geq \frac{(u + v)^2(\sigma_1^2 + \sigma_2^2)}{(\mu_1 - \mu_2)^2}$$

Single proportion

When conducting a particular experiment an event A occurs with an unknown probability π.

With n independent repetitions of the experiment event A occurred x times.

Then, x is an observation of

$$X \sim Bin(n, \pi).$$

We have the estimate $\hat{\pi} = \frac{x}{n}$, that is an observation from the r.v.

$$\frac{X}{n} \approx N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

if $n\pi(1-\pi) > 10$.

We want to test $H_0: \pi = \pi_0$ against $H_1: \pi \neq \pi_0$ with the test on the level α such that that the power for $\pi = \pi_1$ is at least $1 - \beta$.

Test statistics

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

The r.v. Z is approx $N(0,1)$ under H_0 (if $n\pi_0(1-\pi_0) > 10$) and H_0 is rejected if $z < -v$ or $z > v$.

Similar as above we have

Case 3 on "Kirkwoods list":

$$n \geq \frac{\left[u\sqrt{\pi_1(1-\pi_1)} + v\sqrt{\pi_0(1-\pi_0)}\right]^2}{(\pi_1 - \pi_0)^2}$$

Here also $n\pi_1(1-\pi_1) > 10$.

Comparison of two proportions

Assume we have observation x from $Bin(n, \pi_1)$ and y from $Bin(n, \pi_2)$. We want to test

$$H_0 : \pi_1 = \pi_2 \quad \text{against} \quad H_1 : \pi_1 \neq \pi_2,$$

on the significance level α with the test with power at least $1 - \beta$ for a certain difference $\pi_1 - \pi_2$.

Test statistics

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2 - 0}{\sqrt{2\frac{\hat{\pi}(1-\hat{\pi})}{n}}},$$

where $\hat{\pi} = \dfrac{x+y}{2n}$. The r.v. Z is approximately $N(0,1)$ under H_0.

By the similar reasoning as before

Case 6 on "Kirkwoods list":

$$n \geq \frac{\left[u\sqrt{\pi_1(1-\pi_1)+\pi_2(1-\pi_2)}+v\sqrt{2\bar{\pi}(1-\bar{\pi})}\right]^2}{(\pi_2-\pi_1)^2},$$

where $\bar{\pi} = \dfrac{\pi_1+\pi_2}{2}$.

Here, we should have both

$$n\pi_1(1-\pi_1) > 10 \quad \text{and} \quad n\pi_2(1-\pi_2) > 10.$$

Note that one needs to assume both π_1 and π_2 to calculate n.

Generally: Feel free to take larger n value than the calculated one. Remember that often one use some guessed parameter values.

Remark 1 All cases of "Kirkwoods list" are for **two-sided** tests. For one-sided tests we need to exchange v with v', where $\Phi(v') = 1 - \alpha$, compare example above.

Remark 2 In Minitab there are procedures for the determination of sample size for a lot of cases.

Example 3

A certain part of a machine wear due to friction. As a measure of wear one may use the weight loss during a certain operating time.

One has so far used oil A, but is considering switching to oil B.

Make a comparative study by determining the weight loss

$$x_1, \ldots, x_n \text{ for A}$$

$$y_1, \ldots, y_n \text{ for B}$$

and do the comparison test on level 5%.

Problem: How to choose n.

Size of n depends on the variation of the observations and on what criteria to use to be able to show the difference between A and B.

To get an idea of the variance one often do some preliminary investigation. Obtained weight loss:

u_i:	12.7	19.9	13.6	18.3	14.3	20.4

with

$$\bar{u} = 16.53, \quad s_u = 3.40.$$

We assume that r.v. $U_i \sim N(\mu, \sigma)$ and independent.

Having such a small sample, we have a large uncertainty in the σ-estimate s_u.

It is reasonable to make a confidence interval for σ.

We have the r.v. $\frac{5S_u^2}{\sigma^2} \sim \chi^2(5)$, which gives

$$P\left(\frac{5S_u^2}{\sigma^2} > 1.61\right) = 0.90 \quad \Leftrightarrow \quad P\left(\sigma^2 < \frac{5S_u^2}{1.61}\right) = 0.90$$

and confidence interval $I_\sigma = \left(0,\ s_u\sqrt{5/1.61}\right) = (0,\ 6.0)$.

Hence, it is **not** very probable that $\sigma > 6.0$.

We now turn back to the actual survey.

Model: We have two independent random samples from $N(\mu_A,\ \sigma)$ and $N(\mu_B,\ \sigma)$, respectively.

We want to test

$$H_0:\ \mu_A = \mu_B \quad H_1:\ \mu_A \neq \mu_B,$$

on level $\alpha = 0.05 =$ **risk for the error of type I** with a test such that the power

$$P(H_0 \text{ is rejected if } |\mu_A - \mu_B| \geq a) \geq 0.90$$

i.e., such that **risk for the error of type II**

$$P(H_0 \text{ is not rejected if } |\mu_A - \mu_B| \geq a) \leq 0.10 \qquad (1)$$

Point estimators: $\hat{\mu}_A = \bar{x}$; $\hat{\mu}_B = \bar{y}$ together with

$$s^2 = \frac{(n-1)s_A^2 + (n-1)s_B^2}{n-1+n-1}.$$

We obtain $\hat{\mu}_A - \hat{\mu}_B = \bar{x} - \bar{y}$ with $E\left(\overline{X} - \overline{Y}\right) = \mu_A - \mu_B$ and

$$Var\left(\overline{X} - \overline{Y}\right) = Var\left(\overline{X}\right) + Var\left(\overline{Y}\right) = \frac{\sigma^2}{n} + \frac{\sigma^2}{n}.$$

Then it follows that the r.v.

$$\overline{X} - \overline{Y} \sim N\left(\mu_A - \mu_B,\ \sqrt{\frac{2\sigma^2}{n}}\right).$$

Under the null hypothesis $\mu_A - \mu_B = 0$, so we get the statistics

$$w = \frac{\bar{x} - \bar{y} - 0}{s\sqrt{\frac{2}{n}}}.$$

The r.v. $W \sim t(2(n-1))$ if H_0 is true.

H_0 is rejected if $|w| > t$, where t comes from t-table.

In order to take into account the risk of error of type II (1) above, we need

$$d = \left| \frac{\mu_A - \mu_B}{2\sigma} \right| \geq \frac{a}{2\sigma}.$$

Example 4

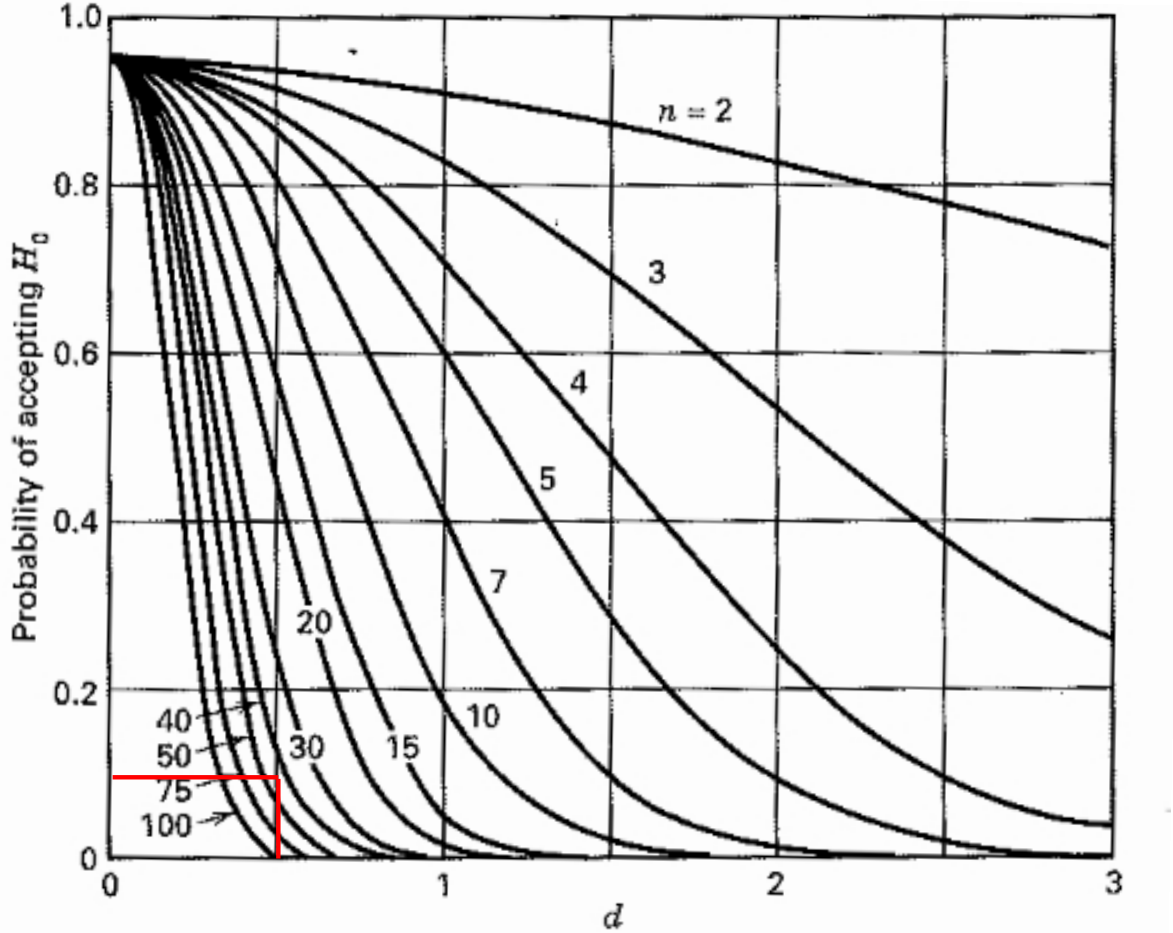
Choose $a = 6$ and we choose $\hat{\sigma} = s = 6$.

We obtain $d \geq \frac{1}{2}$.

From *Operating Characteristic Curves* (OC-curves) we have that

$$n^* = 2n - 1 \approx 45 \quad \text{i.e.,} \quad n \approx 23.$$

Compare with Minitab output below.



■ **FIGURE 2.12** Operating characteristic curves for the two-sided t-test with $\alpha = 0.05$. (Reproduced with permission from "Operating Characteristics for the Common Statistical Tests of Significance," C. L. Ferris, F. E. Grubbs, and C. L. Weaver, *Annals of Mathematical Statistics*, June 1946.)

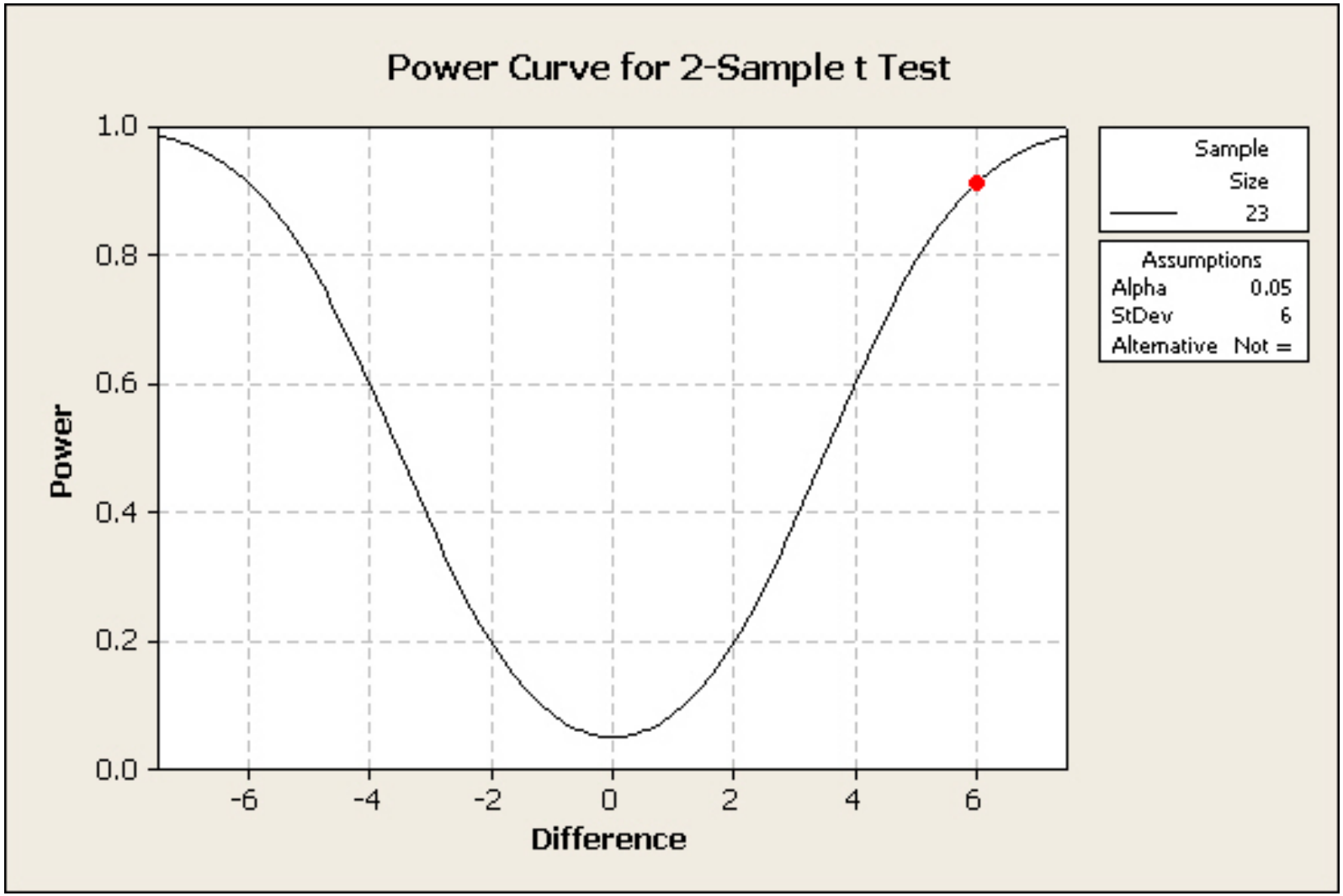
Power and Sample Size

```
2-Sample t Test

Testing mean 1 = mean 2 (versus not =)
Calculating power for mean 1 = mean 2 + difference
Alpha = 0.05  Assumed standard deviation = 6

            Sample  Target
Difference    Size   Power  Actual Power
         6      23     0.9      0.912498

The sample size is for each group.
```


Power Curve for 2-Sample t Test
Power
1.0
0.8
0.6
0.4
0.2
0.0
-6
-4
-2
0
2
4
6
Difference
Sample Size
23
Assumptions
Alpha 0.05
StDev 6
Alternative Not =

Choice of the sample size for one factor design

We assume that we have a samples of the same size, i.e., that $n_1 = \ldots = n_a = n$. Then we have observations y_{il} such that

$$Y_{il} = \mu_i + \varepsilon_{il},$$

for $l = 1, \ldots, n$ and $i = 1, \ldots, a$.

Case 1: We want to test

$$H_0 : \ \mu_1 = \mu_2 = \ldots = \mu_a \quad \text{vs.} \quad H_1 : \ \text{not all } \mu_i \text{ are the same}$$

on level α with a F-test.

One can show that the risk for the error of type II

$$\begin{aligned}\beta &= P(H_0 \text{ is not rejected if } H_0 \text{ is false}) \\ &= P(V < F_\alpha(a-1, N-a) \text{ if } H_0 \text{ is false})\end{aligned}$$

depends only on the parameters ϕ, ν_1 and ν_2, which in our case are

$$\nu_1 = a - 1, \quad \nu_2 = N - a, \quad \phi^2 = n \sum_1^a \frac{\tau_i^2}{a\sigma^2}.$$

- β as function of ϕ for the different ν_1 and ν_2 can be found in appendix V.
- By imposing requirements on β for given values of τ_i and σ one can decide about n.
- See example in the course book.

Case 2: By fixing the length of $I_{\mu_i - \mu_j}$ one can also determine n.

Repetition – Linear regression

Example 5

We want to study a bacteria sample and see what factors affect it. Observations of growth for a sample of 25 bacteria samples were done

y = measure of the amount of bacteria,

x_1 = conditions ($x_1 = 0$ if "moist", $x_1 = 1$ if "dry"),

x_2 = time in hours,

x_3 = temperature.

	x_1	x_2	x_3	y
1	1	73	14	0.5
2	1	66	16	0.5
3	0	65	15	0.7
⋮	⋮	⋮	⋮	⋮
25	0	80	6	3.7

Problem:

(i) How to find the "line" best fit to the points?

(ii) Would a new series of bacteria samples give similar "line"?

(iii) How do we describe the deviations from the "line"?

We answer question (iii) through a model that consider deviations from the "line" as random variables.

Analyze the data using the model

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

where $\varepsilon \sim N(0, \sigma)$.

General regression model

In the general case one would explain the variation of a response variable y using variations of the explanatory variables $x_1, \ldots, x_k$.

One formulates the "theoretical" linear relation

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

To estimate β-coefficients in the linear equation we need to have observed values

$$(x_{11}, x_{12}, \ldots, x_{1k}, y_1), \ \ldots, \ (x_{n1}, x_{n2}, \ldots, x_{nk}, y_n).$$

In general, one discovers then that the relationship between the observed y- and x-values are not exactly linear.

Hence, one need a model which includes deviations from the linear relationship as random variables.

Model

Each y_j is observation of

$$\begin{aligned} Y_j &= \mu_j + \varepsilon_j \\ &= \beta_0 + \beta_1 x_{j1} + \ldots + \beta_k x_{jk} + \varepsilon_j, \end{aligned}$$

for $j = 1, \ldots, n$, where

$$\mu_j = \beta_0 + \beta_1 x_{j1} + \ldots + \beta_k x_{jk}$$

and $x_{j1}, \ldots, x_{jk}$ are fixed known numbers, $\beta_0, \ldots, \beta_k$ are unknown parameters and where

$$\varepsilon_1, \ldots, \varepsilon_n \text{ are independent and } N(0, \sigma).$$

The model gives us the r.v. $Y_1, \ldots, Y_n$ are independent and such as

$$\begin{aligned} Y_j &\sim N(\mu_j, \sigma) \\ &\sim N(\beta_0 + \beta_1 x_{j1} + \ldots + \beta_k x_{jk}, \sigma) \end{aligned}$$

We write the model with matrices instead

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{=\boldsymbol{Y}} = \underbrace{\begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}}_{=\boldsymbol{X}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{=\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{=\boldsymbol{\varepsilon}}$$

or shorter

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Now we have random vector

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\boldsymbol{Y} : n \times 1, \quad \boldsymbol{X} : n \times (k+1), \quad \boldsymbol{\beta} : (k+1) \times 1,$$

with

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{C}_{\boldsymbol{Y}} = \boldsymbol{C}_{\boldsymbol{\varepsilon}} = \sigma^2 \boldsymbol{I},$$

because

$$\varepsilon_1, \ldots, \varepsilon_n \text{ are independent with } \operatorname{var}(\varepsilon_i) = \sigma^2.$$

ML-estimators

Theorem. Under the conditions given in the above model and if $det\left(\boldsymbol{X}'\boldsymbol{X}\right) \neq 0$, then maximum likelihood (ML) estimator is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Furthermore, the random vector

$$\hat{\boldsymbol{\beta}} \sim N_{k+1}\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).$$

Note that

$$\widehat{\mathrm{E}(\boldsymbol{Y})} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

ANOVA

Let $\hat{\mu}_j$ be estimator of $\mathrm{E}(Y_j)$. i.e.,

$$\hat{\mu}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{j1} + \ldots + \hat{\beta}_k x_{jk}.$$

Then, one can show that

$$\underbrace{\sum_{i=1}^{n}(y_j - \overline{y})^2}_{=SS_{TOT}} = \underbrace{\sum_{i=1}^{n}(y_j - \hat{\mu}_j)^2}_{=SS_E} + \underbrace{\sum_{i=1}^{n}(\hat{\mu}_j - \overline{y})^2}_{=SS_{REGR}}.$$

We can also write SS_E as

$$SS_E = \sum_{i=1}^{n}(y_j - \hat{\mu}_j)^2 = \boldsymbol{y}' \left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{y}.$$

Sums of squares can be explained as

- SS_{TOT} the total variation of y-values.
- SS_{REGR} the variation of y-values that are explained by $x_1, ..., x_k$. SS_{REGR} is a quadratic form in $\hat{\beta}_1, ..., \hat{\beta}_k$.
- SS_E the variation of y-values that cannot be explained by the model.

Moreover, we have *the coefficient of determination*

$$R^2 = \frac{SS_{REGR}}{SS_{TOT}}.$$

Theorem. Under the conditions of the above model it holds that

1. The r.v.
$$\frac{SS_E}{\sigma^2} \sim \chi^2(n-k-1).$$
2. The r.v. SS_E is independent of the r.v. SS_{REGR} and $\hat{\boldsymbol{\beta}}$.
3. If $\beta_1 = \beta_2 = ... = \beta_k = 0$ then the r.v.
$$\frac{SS_{REGR}}{\sigma^2} \sim \chi^2(k).$$

The theorem above gives

1. We have an unbiased variance estimate

$$s^2 = \frac{SS_E}{n-k-1}.$$

2. Independence can be used in the construction of confidence intervals.
3. Test the hypothesis

$$H_0: \ \beta_1 = ... = \beta_k = 0 \quad \text{against} \quad H_1: \ \text{at least one } \beta_i \neq 0.$$

using the statistic

$$V = \frac{SS_{REGR}/k}{SS_E/(n-k-1)} \underset{H_0}{\sim} F(k, n-k-1)$$

and H_0 is rejected for large values of v.

Parameter estimators - Properties

If

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} h_{00} & h_{01} & \dots & h_{0k} \\ h_{10} & h_{11} & \dots & h_{1k} \\ \vdots & \vdots & & \vdots \\ h_{k0} & h_{k1} & \dots & h_{kk} \end{pmatrix}$$

we have $\hat{\beta}_i \sim N(\beta_i, \sigma\sqrt{h_{ii}})$. Hence,

1. $$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{h_{ii}}} \sim N(0,1)$$

2. $$\frac{(n-k-1)s^2}{\sigma^2} \sim \chi^2(n-k-1)$$

3. the r.v. $\hat{\beta}_i$ and s^2 are independent.

This now gives our help variable to construct confidence intervals

$$\frac{\dfrac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{h_{ii}}}}{\sqrt{\dfrac{s^2}{\sigma^2}}} = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{h_{ii}}} \sim t(n-k-1),$$

which can be used for hypothesis testing or confidence interval.

Generally - Confidence interval for $E(Y_0)$

We are interested in a new yet not observed r.v.

$$Y_0 = \underbrace{\beta_0 + u_1\beta_1 + \ldots + u_k\beta_k}_{=\mu_0} + \varepsilon_0 = \mu_0 + \varepsilon_0,$$

where $u_1, \ldots, u_k$ are the current value of the explanatory variables, i.e. known number, where ε_0 is independent of $\varepsilon_1, \ldots, \varepsilon_n$ and

$$\varepsilon_0 \sim N(0, \sigma).$$

We introduce the vector

$$\boldsymbol{u}' = (1 \; u_1 \; \ldots \; u_k)$$

and we can write

$$Y_0 = \boldsymbol{u}'\boldsymbol{\beta} + \varepsilon_0.$$

Now we want to estimate $\mu_0 = \boldsymbol{u}'\boldsymbol{\beta}$ and we do it with $\boldsymbol{u}'\widehat{\boldsymbol{\beta}}$

The estimate is an observation from the r.v. $\hat{\mu}_0 = \boldsymbol{u}'\widehat{\boldsymbol{\beta}}$.

Expected value of $\hat{\mu}_0$ is

$$\mathrm{E}(\hat{\mu}_0) = \mathrm{E}\left(\boldsymbol{u}'\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{u}'\,\mathrm{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{u}'\boldsymbol{\beta} = \mu_0$$

and variance is given by

$$\mathrm{var}(\hat{\mu}_0) = \mathrm{var}\left(\boldsymbol{u}'\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{u}'\boldsymbol{C}_{\boldsymbol{\beta}}\boldsymbol{u} = \boldsymbol{u}'\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}.$$

Moreover, as a r.v. $\hat{\mu}_0$ is a linear transformation of a Gaussian vector $\Rightarrow$ $\hat{\mu}_0$ becomes normally distributed, i.e.,

$$\hat{\mu}_0 = \boldsymbol{u}'\widehat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{u}'\boldsymbol{\beta}, \sigma\sqrt{\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}}\right).$$

Help variable to construct confidence intervals for $E(Y_0) = \mu_0 = \boldsymbol{u}'\boldsymbol{\beta}$ is

$$\frac{\hat{\mu}_0 - \mu_0}{S\sqrt{\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}}} = \frac{\boldsymbol{u}'\widehat{\boldsymbol{\beta}} - \boldsymbol{u}'\boldsymbol{\beta}}{S\sqrt{\boldsymbol{u}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{u}}} \sim t(n-k-1).$$

Observe that degrees of freedom in t-distribution always comes from σ^2-estimator of χ^2-variable.

Note that the expectation μ_0 is a "theoretical" value that describes what one obtain in the average if one do many measurements with the same value of explanatory variables.

Model choice in regression analyze

When studying the response variable Y, one often choose between number of the alternative models, through the different choice of the explanatory variables. The aim is to explain as much of the variation in the y values as possible by utilizing relevant explanatory variables.

Models are compared using the residual sums, or residual mean sums of squares (σ^2-estimates).

Small value of σ suggests that ε-variables in the model is quite negligible.

Residual analysis

The residuals are defined as

$$e_j = y_j - \hat{\mu}_j = y_j - \hat{\beta}_0 - \hat{\beta}_1 x_{j1} - \cdots - \hat{\beta}_k x_{jk}.$$

Hence, this what approximates ε-variables.

Using residual plots one can for example detect deviations from linearity or Gaussian distribution.

Dummy variables

To separate different categories within the model one can utilize so-called *dummy variables* (indicator variables). For p categories one needd $p - 1$ dummy variables.

One level will always be some sort of ground state, see Ex. below.

Example 5, cont.

We want to study a bacteria culture and see what factors affect it. Measurements for growth for a sample of 25 bacteria cultures were done

$$
\begin{aligned}
y &= \text{measure of the amount of bacteria,} \\
x_1 &= \text{conditions } (x_1 = 0 \text{ if "moist", } x_1 = 1 \text{ if "dry"}), \\
x_2 &= \text{time in hours,} \\
x_3 &= \text{temperature.}
\end{aligned}
$$

a) Which model is the data analyzed with?

b) Is the relationship significant?

c) How much bacteria we have after 75 hours in 15 degrees?

Regression Analysis: y versus x1, x2, x3

```
The regression equation is
y = 4.54 - 1.11 x1 + 0.0016 x2 - 0.184 x3

Predictor      Coef  SE Coef      T      P
Constant     4.5403   0.8374   5.42  0.000
x1          -1.1068   0.3268  -3.39  0.003
x2          0.00160  0.01038   0.15  0.879
x3         -0.18415  0.04435  -4.15  0.000

S = 0.779189   R-Sq = 60.0%   R-Sq(adj) = 54.3%

Analysis of Variance

Source           DF       SS      MS      F      P
Regression        3  19.1102  6.3701  10.49  0.000
Residual Error   21  12.7498  0.6071
Total            24  31.8600
```

Linköping University - Research that makes a difference