Experimental Design and Data Analysis, Lecture 7

Eduard Belitser

VU Amsterdam

Lecture overview

1. contingency tables
    1. chisquare test
    2. Fisher test
2. simple linear regression
3. multiple linear regression

contingency tables

Setting

An experiment with:

- a count of individuals or units in different categories of two factors.

should be smth that is conn. to cert. fac.

you divide pop. into levels of fac. = groups acc. to 2 cat.s

you have cells = comb. of levels b.w. 2 fac.s

Interest is in a possible dependence of the two factors.

EXAMPLE Study possible dependency between blood group and disease by counting the number of patients having a certain blood group (A, B or O) and a certain disease (stomach cancer, kidney cancer, no disease).

everyone in pop. is in 1 comb. of levels

EXAMPLE Study possible dependency between web layout and size of a company by counting the number of companies of a certain size (small, moderate, large) using a certain web design (relative, fixed, elastic, liquid) .

EXAMPLE Consider the following (fictive) counts amongst 60 VU-students:

	exact	arts	total
men	23	17	40
women	7	13	20
total	30	30	60

Question: study and gender independent?

Design

draw ppl. from pop. rand.ly

Design A:

- Take a random sample of experimental units from the relevant population.
- Count for each cross-category the number of units falling into that cross-category.

fix one cat. → take rand. samp. of experiment. units from each cat.

Design B:

- Take for each category of the first (row) factor a random sample of experimental units.
- Count for each category of the second factor the number of units falling into that cross-category.

same as design B but we exchange cat.s

Design C:

- Take for each category of the second (column) factor a random sample of experimental units.
- Count for each category of the first factor the number of units falling into that cross-category.

Analysis (1)

"." : summat. over 2nd index → computing # of ppl in 1st level of 1st fac.

The general form of a contingency table is

n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
$\vdots$		$\ddots$	$\vdots$	$\vdots$
n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	$n_{\cdot\cdot}$

IxJ metrics/ num.s

⊙ 1st fac. = row fac.
⊙ 2nd fac. = col. fac.

total num. of obs./ppl.

We want to test whether the two factors are independent (under design A):

H_0 : *row variable and column variable are independent.*

dist. that runs how pop. dist. over levels of 1st fac. is indep. of dist. how ppl. are dist acc. to 2nd fac.

Or, we want to test whether the distributions are homogeneous over rows (design B) or columns (design C):

H_0 : *the distributions over row (column) factors are equal.*

dist. over rows = num.s can be diff. but frac.s are same / they are propart. = homogenity w.r.t. fac.s

Analysis (2)

Let $n = n_{\cdot\cdot}$ be the total number of observaions. Under the null hypothesis of no dependence (or homogeneity), the counts are expected to be in proportion:

$$E_{ij} = np_{ij} = np_{i\cdot}p_{\cdot j} = n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n} = \frac{n_{i\cdot}n_{\cdot j}}{n}.$$

no dep. b.w. fac.s

E_{ij} → exp. num. of indiv.s in cell

$\frac{n_{i\cdot}n_{\cdot j}}{n}$ → can be computed

* P_{ij} = true prob. of cell → unknown

Expected counts in the example data set:

	exact	arts	total
men	?	?	40
women	?	?	20
total	30	30	60

$\Longrightarrow$

	exact	arts	total
men	$60 \cdot \frac{40}{60} \cdot \frac{30}{60}$	$60 \cdot \frac{40}{60} \cdot \frac{30}{60}$	40
women	$60 \cdot \frac{20}{60} \cdot \frac{30}{60}$	$60 \cdot \frac{20}{60} \cdot \frac{30}{60}$	20
total	30	30	60

2x2 table

The test statistic is based on the (appropriately normalized) differences between the expected counts E_{ij} under H_0 and the observed counts n_{ij}:

$$T = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)},$$

chi-sq. dist. ($I, J \geq 2$)

(approx. a chisquare distribution).

we div. by E_{ij} for normaliz. o.w. numerat. will be v. big if samp. size is big.

expect. val. = if H_0 is true, then our obs. should be close to this num.

The p-value is always right-sided: $p_{right} = P(T > t)$. Why?

because $T > t$ means we deviate hypoth. significantly

Condition: For the test to be reliable, at least 80% of the E_{ij}'s should be at least 5.

for chi-sq. to perf. well

In R: `chisq.test(data)` → matrix

Analysis in R: data input

First, we need to create a table of the counts in the form of a matrix.

The following data consists of grade counts in an elementary statistics class, classified by the students' majors.

```
> grades=matrix(c(8,15,13,14,19,15,15,4,7,3,1,4),byrow=TRUE,ncol=3,nrow=4,
+ dimnames=list(c("A","B","C","D-F"),c("Psychology","Biology","Other")))
> grades
```

	Psychology	Biology	Other
A	8	15	13
B	14	19	15
C	15	4	7
D-F	3	1	4

# of col.s & rows

assign names of col.s and rows

grades

students by major

⊙ we want to know whether dist. of grades is same for all studies or students of some major is doing better at stat. than others

For the calculations on the next slide, *R* needs the data in a `matrix` object, rather than in a `table` or `dataframe` format.

Analysis in R: testing (1)

```
> rowsums=apply(grades,1,sum); colsums=apply(grades,2,sum)
> total=sum(grades); expected=(rowsums%*%t(colsums))/total
> round(expected,0)
     Psychology Biology Other
[1,]         12      12    12
[2,]         16      16    16
[3,]          9       9     9
[4,]          3       3     3
> sum((grades-expected)^2/expected) #realization of statistics T
[1] 12.18346
> 1-pchisq(12.18346,6)   #p-value for the observed T=12.18346
[1] 0.05799897
```

rounded to get int. val.s

expect. val.s

if our H_0 is true, then the val.s should be like this

realiz. of chi-sq. dist. r.v. w/ 6 d.o.f.

d.o.f.

> 0.05

prob. of chi-sq. dist. r.v. w/ 6 d.o.f bigger than that num.

Less than 80% of the expected counts are above 5. Hence, the approximation by a chi-square test is not reliable.

only 75% of exp. counts are above 5.

Analysis in R: testing (2)

Of course, no need to perform all these computations, just use build-in R command: `chisq.test`, which executes the χ^2-test.

```
> z=chisq.test(grades); z
                    Pearson's Chi-squared test

data:  grades
X-squared = 12.1835, df = 6, p-value = 0.058

Warning message:
In chisq.test(grades) : Chi-squared approximation may be incorrect
```

H_0 is not rejected.

b/c less than 80% of exp. counts are above 5 = unreliable test

R gives a warning because the chi-squared approximation in this case is not reliable. In such a case one can use the setting `simulate.p.value=TRUE`, which computes a *p*-value in a bootstrap fashion. This may yield a very different *p*-value.

```
> chisq.test(grades,simulate.p.value=TRUE)
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  grades
X-squared = 12.1835, df = NA, p-value = 0.05647
```

Analysis in R: testing (3)

* if p-val. is small, fac.s are indep.= they don't interact w/each other → we would like to see which cell contrib. more to the diff.

You can extract information from z=chisq.test(grades): z$expected gives the table of expected values, z$observed recovers the observed values. We can look at the (square root) contributions of each cell to the chi-squared statistics, by using residuals(z) (or z$residuals), to determine which observed values deviate most from the expected under H_0.

```
> residuals(z)  # = (z$observed-z$expected)/sqrt(z$expected)
     Psychology    Biology      Other
A    -1.2032599  0.8992005  0.3193881
B    -0.5630451  0.7872412 -0.2170232
C     2.0838439 -1.5668929 -0.5434979
D-F   0.1749697 -1.0110751  0.8338764
```

sq. root: contribution of each cell = we keep the sign b/c we want to know if n_{ij} is bigger than E_{ij}

- From this table we see that psychology students have relatively more C's, → biggest val.
- biology students have relatively less C's, → 2nd biggest val.
- psychology students have relatively less A's, → 3rd biggest val. → psy. students aren't doing well w.r.t. high grades

than expected under H_0 (the differences are not significant though ($p \approx 0.06$)).

Alternatively, we can look at the standardized residuals using the command z$stdres (=(z$observed-z$expected)/sqrt(V), where V is the residual cell variance, see Agresti, 2007, section 2.4.5) and compare this to $z_{\alpha/2}$=qnorm(0.975)≈1.96. → has less mean. to interpret

Fisher's exact test for 2x2-tables

* 2x2 tables = special case of contin. tables → 2 cat.s (yes/no etc.)

For 2x2-tables it is possible to compute an exact p-value, that does not use approximation or simulation. This is called Fisher's exact test.

→ chi-sq. test uses approx

Data on right- and left-handed people, classified according to gender.

```
> handed=matrix(c(2780,3281,311,300),nrow=2,ncol=2,byrow=TRUE,
+ dimnames=list(c("right-handed","other"),c("men","women")))
> handed
```

2 levels

	men	women
right-handed	2780	3281
left-handed	311	300

2 levels

right-hand.

* n_{11} = has a cert. prob. → if it's too small, then H_0 is not true b/c we would have got smth more probable

We can compare this to picking without replacement 3091 balls from a vase which contains 6672 balls, 6061 white and 611 red. The number of white balls amongst the picked 3091 balls is $n_{11} = 2780$.

left-hand.

n_{11}	...	6061
...	...	611
3091	3581	6672

$\Longrightarrow$

n_{11}	$6061 - n_{11}$
$3091 - n_{11}$	$3581 - (6061 - n_{11})$

The number n_{11} determines all other numbers. Fisher's exact test is based on this number. Under the null hypothesis of no dependence between the two factors it has a hypergeometric distribution.

Analysis in R: testing

```
> fisher.test(handed)

        Fisher's Exact Test for Count Data

data:  handed
p-value = 0.01918
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6894895 0.9688105
sample estimates:
odds ratio
 0.8173619
> chisq.test(handed)

        Pearson's Chi-squared test with Yates' continuity correction

data:  handed
X-squared = 5.4542, df = 1, p-value = 0.01952
```

computes prob. of hypergeo. dist. w/given param.s and the realization

unprobable realization = not true that they are indep.

being M/W or L/R actually matters.

if it's smaller than 1, it gives you the direction of deviation = there are more left-handed men.

The chisquare approximation is also fine for these data. The odds ratio is computed as $\frac{2780/311}{3281/300} = 0.8173619$ and can be interpreted as "for one right-handed women there is ≈ 0.82 right-handed men", there are relatively more left handed men than women.

simple linear regression

Setting

An experiment with:

- a numerical outcome Y ("dependent variable"),
- a numerical explanatory variable X ("independent variable").

The purpose is to explain Y by a numerical function of X. Extrapolation to nonmeasured values of X is desirable.

EXAMPLE Chemical production process with outcome total yield and explanatory variable temperature.

EXAMPLE Educational study with outcome score on final exam and explanatory variable number of pupils per teacher.

EXAMPLE Quality of a genetic algorithm to determine the minimal value of a criterion function with outcome CPU time needed to find true minimum and explanatory variable mutation probability.

Design

- Fix a set of values X of the explanatory variable.
- Perform the corresponding experiments and measure the outcome Y.

It is natural to let the explanatory variable X vary over a grid of values in its range of interest.

Regression analysis is also often used in nonexperimental situations, with the explanatory variable not under control.

Analysis

Data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

* noise → norm dist. w/ mean = 0, var. = σ^2 and indep.

The simple linear regression model assumes that

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad i = 1, 2, \ldots, n, \quad e_1, \ldots, e_n \sim N(0, \sigma^2).$$

linear func. — noise

We test the null hypothesis $H_0 : \beta_1 = 0$ that the explanatory variable does *not* influence the outcome. We also want to estimate the parameters β_0, β_1.

there is no conn. b.w. X and Y → X has no effect on Y

intercept — slope

The function $x \mapsto \beta_0 + \beta_1 x$ is a line with intercept (value at $x = 0$) β_0 and slope (change per unit) β_1. This is a simple function and may give a bad fit!

Analysis in R: data input, graphics, estimation and testing

The column `total` of the dataset `sat.txt` is the average score on the *scolastic aptitude test* of pupils in US states in 1994/95; the column `expend` is the amount of dollars spent per pupil in the state.

```
> sat=read.table("sat.txt",header=TRUE); sat1=sat[,c(1,7)]; sat1[1:2,]
           expend total
Alabama     4.405  1029
Alaska      8.963   934
```

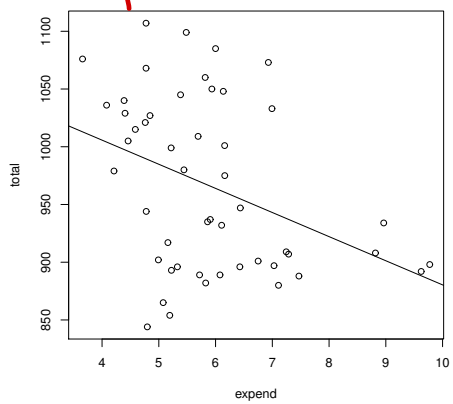
```
> sat1lm=lm(total~expend,data=sat1); summary(sat1lm)
[ some output deleted ]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1089.294     44.390  24.539  < 2e-16 ***
expend       -20.892      7.328  -2.851  0.00641 **
```

you would expect the more money state spends, students perf. better = we see that's not the case when we fit lin. regr.

The parameters β_0 and β_1 are estimated to be 1089.294 and -20.892. The p-value for testing $H_0 : \beta_1 = 0$ is 0.00641. The slope is significantly negative!

→ p-val. is small

```
> plot(total~expend,data=sat1)
> abline(sat1lm)
```



Compare to Pearson's correlation test

Compare simple linear regression to Pearson's correlation test (treated earlier) which tests whether the response and explanatory variable (in our case columns total and expand) are uncorreleted.

```
> cor.test(sat1$total,sat1$expend)

Pearson's product-moment correlation

data:  sat1$total and sat1$expend
t = -2.8509, df = 48, p-value = 0.006408
```

we got the same num.

Notice that the p-value of the correlation test between response and covariate is equal to the p-value for testing the zero slope in simple linear regression. In fact, this is the same test: testing $H_0 : \rho = 0$ is the same as testing $H_0 : \beta_1 = 0$.

PCT SLR

Analysis in R: diagnostics

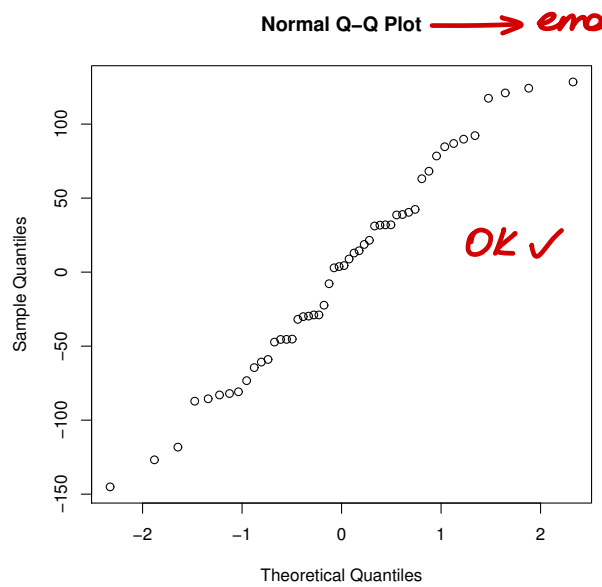
We can use the data to check whether the assumptions on the errors $e_i = Y_i - \beta_0 - \beta_1 X_i$ are not totally untrue.

The residuals are $\hat{e}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$; the fitted values $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. The residuals should look normal, and their spread should not vary with the fitted values.

est. response

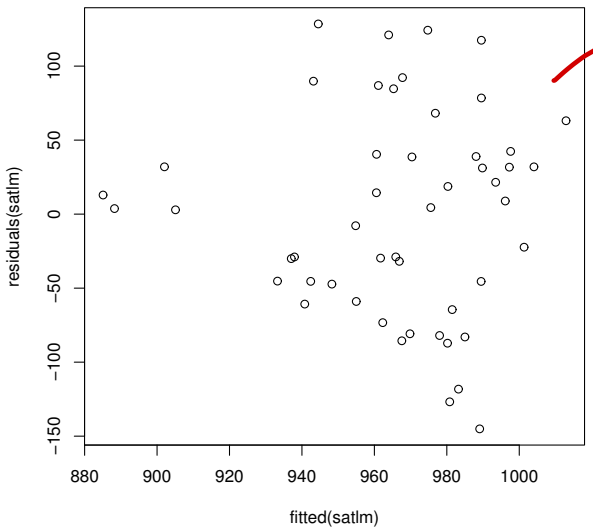
est. of error

```
> qqnorm(residuals(sat1lm))
> plot(fitted(sat1lm),residuals(sat1lm))
```



error is supposed to be normal.

OK ✓



OK ✓

The two plots look ok.

multiple linear regression

Setting and design

Setting: an experiment with

- a numerical outcome Y ("dependent variable");
- p numerical explanatory variables $X_1, \ldots, X_p$ ("independent variables", "predictors").

The purpose is to explain Y by a numerical function of $X_1, \ldots, X_p$.

EXAMPLE Chemical production process with outcome total yield and explanatory variables temperature and pressure.

EXAMPLE Educational study with outcome score on final exam and explanatory variables teacher salaries and number of pupils per teacher.

for each obs. you measure temp. and pressure → then you get a yield for that

Design:

- Fix a set of combinations $(X_1, \ldots, X_p)$ of explanatory variables.
- Perform the corresponding experiments and measure the outcome Y.

we can choose val.s or they can be given e.g. someone's weight = given, time moments = chosen

It is natural to let each explanatory variable vary over a grid and use all their possible combinations, but this may necessitate many experiments. (Regression analysis is also often used in non-experimental situations, with the explanatory variables not under control.)

Analysis

double indexing of X var.

$X_{i1}, \ldots, X_{ip}$ = each of them is measured for ith obs.

Data $Y_i, X_{i1}, X_{i2}, \ldots, X_{ip}$, $i = 1, \ldots, n$. The linear regression model:

there are p var. X and

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + e_i, \quad i = 1, \ldots, n, \quad (\text{matrix notation } Y = X\beta + e)$$

for each one of them, we measure n obs.

vec. of err.s

where errors $e_1, e_2, \ldots, e_n$ are viewed as a random sample from $N(0, \sigma^2)$, $\beta_0, \ldots, \beta_p$ are unknown population parameters.

We test the null hypotheses $H_0 : \beta_j = 0$ that the jth explanatory variable does *not* influence the outcome for $j = 1, \ldots, p$.

We also want to estimate the parameters β_j's.

we want to see how response is affect. by var.s

Possible explanatory variables (prediction variables):

- all x_i different $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_7 x_7 + e$,
- powers of x_j's $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + e$,
- interactions between x_j's $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$.

⊙ each X_{ip} is a col. vect. of mat. X and each entry in a col. is a measurem. of that var. for particular obs.

Essential: all models are linear in β_j's, but not necessarily in x_j's.

as long as it is linear w.r.t. β_j's, it's a linear model

x_j's are not imp.

All ANOVA models can also be written in the matrix notation $Y = X\beta + e$, for some design matrix X (composed of "dummy variables"), where β is the vector of all the ANOVA coefficients involved. Thus the rest of this part also relates to all ANOVA models.

Estimating parameters, SSE

smallest val. of func. → characterizes fit = smaller the func. better the fit → smallest predict. error w/in model

To find the best parameters we minimize the sum of squared errors:

$$\min_{\beta_0,\ldots\beta_p} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_p X_{ip})^2 = RSS,$$

func. of β's → we can compute b/c Y, X's are known, β's are unknown.

estimates = found by using Y and X's

surrog. for err.

$\hat{\beta}_0, \ldots, \hat{\beta}_p$ are the least squares estimates, RSS is the Residual Sum of Squares.

Notation: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ik}$ is called prediction/predicted response.

The Residual Sum of Squares RSS (also called Sum of Squared Errors, SSE) and the estimated variance of the errors e_n:

$$RSS = SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2, \qquad \hat{\sigma}^2 = s^2 = \frac{RSS}{n-p-1} = \frac{SSE}{n-p-1}.$$

sum of sq. residuals

$\hat{\sigma}^2$ is the estimated variance of the e_i's, $\hat{e}_i = Y_i - \hat{Y}_i$ is the i-th residual (the estimated error e_i of the i-th observation).

In R: `model=lm(y~x1+...+xp,data=...)`

response var.

Coefficient of determination R^2

we want to compare fit of our model w.r.t. b.g. model = if diff. is big then the predictors make diff. = fit matters (var.s)

- The coefficient of determination (also called the proportion of explained variance) R^2 compares the fits for the models

b.g. model w/ no var.s

$\omega: \ Y = \beta_0 + e$ and $\Omega: \ Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + e.$

our model w/ p var.s

- For model ω, $\hat{\beta}_0 = \bar{Y}$, the fit is $SS_y = \sum_{i=1}^n (Y_i - \bar{Y})^2$, called total SS.
- For model Ω, the fit is $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, the residual SS.

if it's v. small, we have v. good fit

- explained variation $= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$. (total variation − unexplained variation)
- The coefficient of determination R^2 is defined as

$$R^2 = \frac{SS_y - RSS}{SS_y} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{explained variation}}{\text{total variation}}.$$

always worse b/c it's based on 1 unknown β_0.

close to 1 = big diff.s
close to 0 = no much contrib. of var.s

- $0 \leq R^2 \leq 1$ because always $SS_y \geq SSE \geq 0$.
- R^2 yields a global check on the multiple linear regression model.
- The higher R^2, the more variation the model explains.
- If $p = 1$, then $R^2 = r^2$ (the squared correlation between X_1 and Y).

★ $R^2 \approx 1$ means that the linear regression model can explain the measured response values Y very well using a linear function of the explanatory variables $(X_1, \ldots, X_p)$.

★ $R^2 \approx 0$ means that the linear model does not explain much.

Global model fit

- Data: $X_{i1}, X_{i2}, \ldots, X_{ip}, Y_i,, \; i = 1, \ldots, n.$
- Assumption: the ind. errors follow a $N(0, \sigma^2)$-distribution.
- When is the linear regression model adequate as a whole? In linear regression we compare the models

we test whether ω is as good as Ω = if that's true, we should not use Ω, it's useless ↓ we want to test it globally

$\omega : Y = \beta_0 + e$ and $\Omega : Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e.$

- Test if $X_1, \ldots, X_p$ together have significant explanatory power in the model: $H_0 : \beta_1 = \ldots = \beta_p = 0$ versus H_1 : at least one $\beta_i \neq 0$.
- The test statistic: under H_0, $T = \frac{R^2/p}{(1-R^2)/(n-(p+1))} \sim F_{p,n-(p+1)}$.
  Notice that the case $p = 1$ corresponds to Pearson's correlation test.
- The larger R^2 (hence T is large), the more evidence against H_0, hence we reject H_0 if T is large.
- The right-sided test: for $T \sim F_{p,n-(p+1)}$, reject H_0 if $p = P(T > t) < \alpha$.
- In R: this p-value is in the last line of `summary(model)`.

Relevance of individual coefficients

- Not all available explanatory variables may have explanatory power.
- From all explanatory variables, we need to find relevant ones by testing for individual coefficients.
- Test $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ for individual β_i's (usually two-sided).
- The setting and assumptions are the same as before.
- Test statistic: under H_0,

$$T_i = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \sim t_{n-(p+1)},$$

where $s^2_{\hat{\beta}_i} = \hat{\sigma}^2 \nu_{ii}$, $[\nu_{ij}] = (X^T X)^{-1}$, $Y = X\beta + e$.

- In R: the estimates $\hat{\beta}_i$, standard errors $s_{\hat{\beta}_i}$, the statistics values T_i and the p-values are (in the column `Pr(>|t|)`) all given in the output of `summary(model)`.

Example: bodyfat data

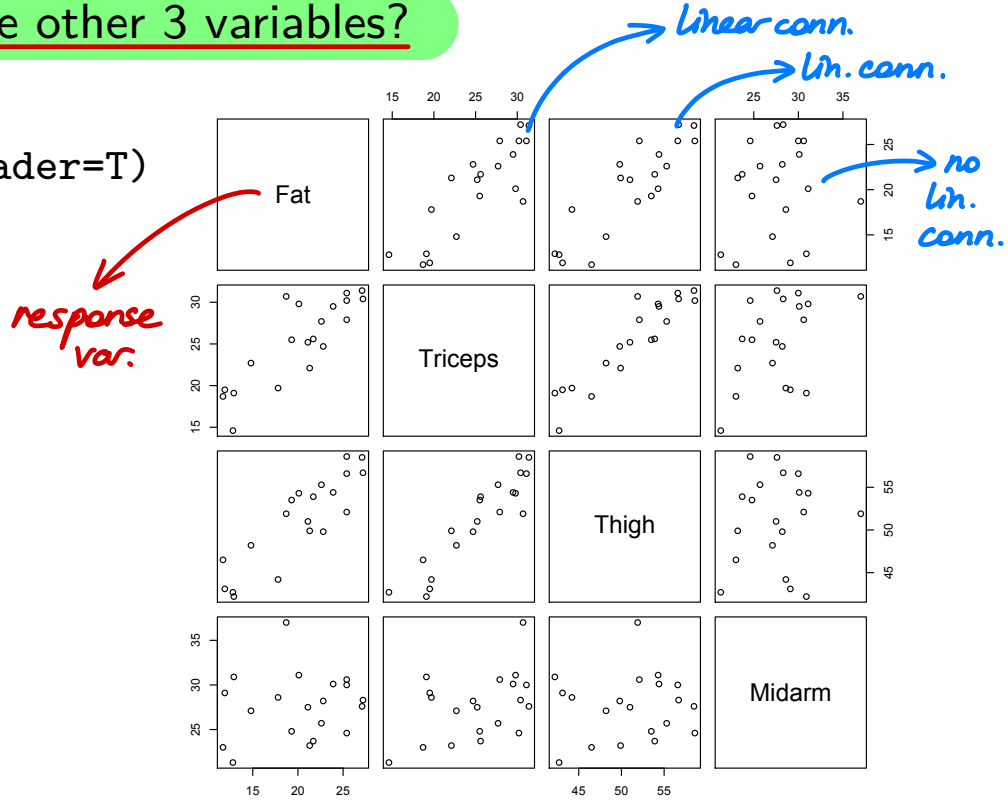
Data of 20 individuals between 25 and 30 years old on amount of body fat, triceps skinfold thickness, thigh circumference and midarm circumference.
Body fat is hard to measure, while the other 3 variables are easy to measure.
Question: can we predict Fat from the other 3 variables?

```
> bodyfat=read.table("bodyfat.txt",header=T)
> bodyfat
   Fat Triceps Thigh Midarm
1  11.9   19.5  43.1   29.1
2  22.8   24.7  49.8   28.2
3  18.7   30.7  51.9   37.0
...
19 14.8   22.7  48.2   27.1
20 21.1   25.2  51.0   27.5
```

X_1 X_2 X_3 Y = Fat

Scatter plots of all pairs:

```
> pairs(bodyfat)
```



Example: bodyfat data

```
> bodyfatlm=lm(Fat~Triceps+Thigh+Midarm,data=bodyfat); summary(bodyfatlm)
 [some output is deleted]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.085     99.782   1.173    0.258
Triceps       4.334      3.016   1.437    0.170
Thigh        -2.857      2.582  -1.106    0.285
Midarm       -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014, Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

Many things can be read from this output. The estimates $\hat{\beta}_i$ are in the column Estimate, $\hat{\sigma} = 2.48$ (so $\hat{\sigma}^2 = 6.15$), $s^2_{\hat{\beta}_i}$'s are in the column Std. Error, T_i's in the column t value, the p-values for individual tests $\beta_i = 0$ are in column Pr(>|t|). The CI's for the β_i's are $\hat{\beta}_i \pm t_{\alpha/2,n-(p+1)} s_{\hat{\beta}_i}$, obtained in R by confint(bodyfatlm). Next, $R^2 = 0.8014$, $R^2_{adj} = 0.7641$. For testing the global model fit, statistics $T = 21.52$, the p-value=7.343e-06. From this output: none of the β_i's is individually significant, but all together they are significant and explain 80%!

Adjusted R^2

if you add enough X's, you get total fit → $R^2 = 1$

not a big deal to have R^2 close to 1 if you have many var.s = we have to penal. for having many var.s

- We want a good fit (high R^2) and a small number of explan. variables.
- Since more explanatory variables always explain more, R^2 always increases with more variables. R^2 can be found in the output of `summary(model)`.
- One considers the R^2 adjusted for the number k of explanatory variables:

$$R^2_{adj} = 1 - \frac{n-1}{n-(p+1)}(1 - R^2).$$

It would not grow at some point like R^2 even after adding more var.s

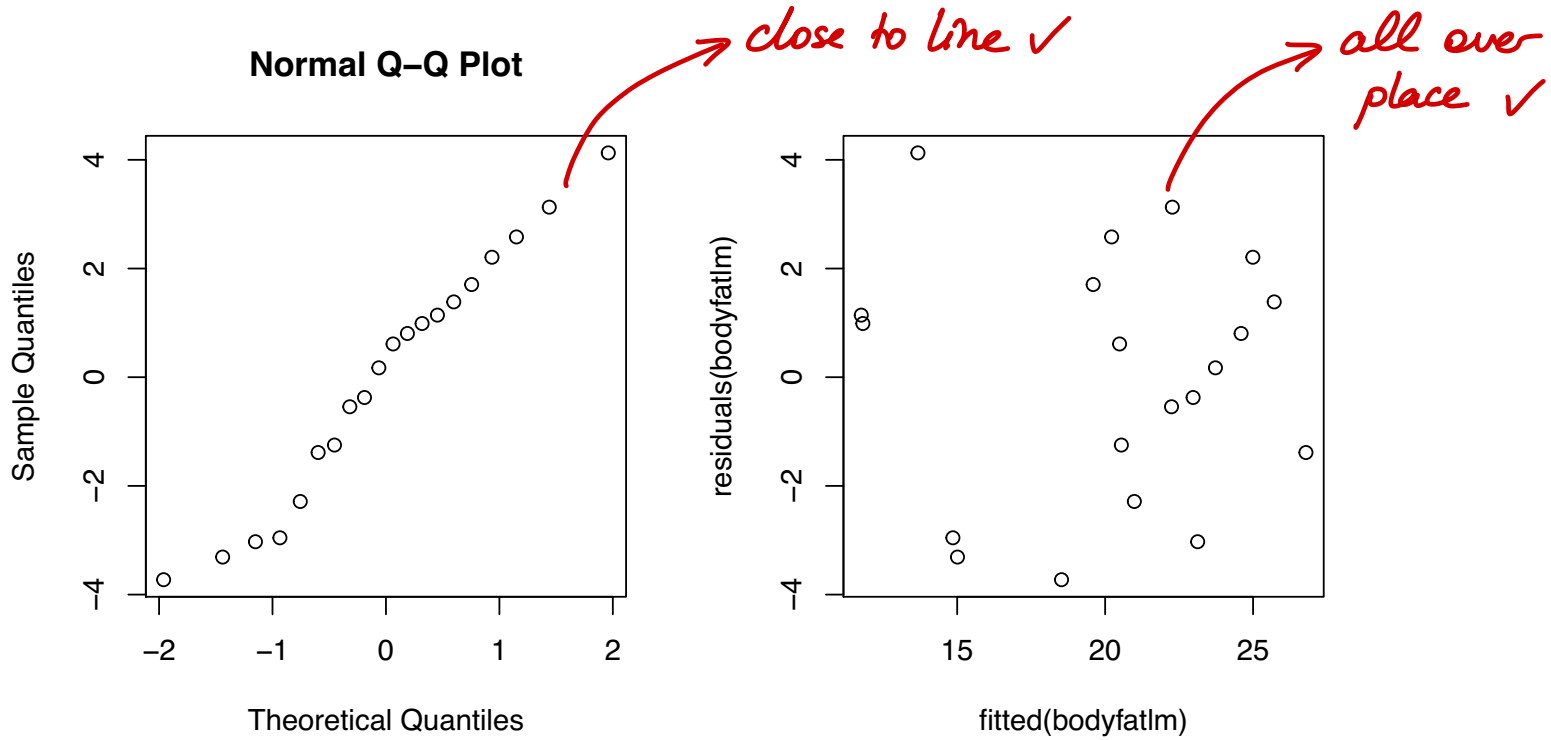
- The more variables, the more conservative R^2_{adj} becomes (as compared to R^2), it can be used to choose between models with different amounts of variables. R^2_{adj} can also be found in the output of `summary(model)`.
- The interpretation of R^2_{adj} is not fraction of explained variance anymore.

doesn't explain anything

Analysis in R: diagnostics

The residuals $\hat{e}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip}$ (in R: `residuals(model)`); the fitted values $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$ (in R: `fitted(model)`).

```
> qqnorm(residuals(bodyfatlm))
> plot(fitted(bodyfatlm),residuals(bodyfatlm))
```



Both plots look ok.

If the assumptions fail?

One can consider:

- transforming the outcomes (e.g., use $\log Y$, Y^3).
- transforming the explanatory variables (e.g. use $\log X$, X^2).
- adding powers $X_i^2, X_i^3, \ldots$ of the regression variables.
- adding "interactions" like $X_i X_j$.
- performing nonparametric or additive regression.
- something else (there is no fix that always works).

To finish

Today we discussed:

- contingency tables
  - chi-square test
  - Fisher test
- simple linear regression
- multiple linear regression

Next time: more on linear regression.