

## Index of 11-L

Page	Title
1	Practical information
2	Data examples
3	Factorial designs
4	Notation for two-way ANOVA
5	One-way ANOVA for two-way factorial
6	Decomposing a two-way table of means I
7	Decomposing a two-way table of means II
8	Additivity and interaction
9	Two-way ANOVA models
10	Analysis of two-way ANOVA (with replication)
11	Summary of two-way ANOVA for tomato data
12	Exercises 13.3 and 13.4
13	Multiple linear regression model
14	Model assumptions and interpretations
15	Multiple linear regression analysis
16	Interpretation of regression coefficients
17	Model building / variable selection
18	Summary of analysis for CSDATA example
19	Model checking for ANOVA and multiple regression
20	Overview one-way & two-way ANOVA
21	Summary notes

## PRACTICAL INFORMATION

### Schedule:

- third **home assignment** to be returned to you on Monday,
- fourth **home assignment** (statistical reporting in papers) underway, due next Thursday.

### Today's lecture — introductions<sup>1</sup> to two-way ANOVA and multiple regression:

- **two-way ANOVA**<sup>2</sup> — first extension beyond single predictor models,
  - \* impact on outcome of two factors (categorical variables) studied simultaneously  
→ new concept of **interaction** (and additivity),<sup>3</sup>
- **multiple regression** = regression with more than one  $x$ -variable,<sup>4</sup>
  - \* model looks similar to simple linear regression, but
    - different interpretation of parameters,
    - new issue: **selection** of  $x$ -variables,
- both models fall within framework of (general) **linear models** with many aspects of analysis similar to what we have seen: least square estimation, confidence intervals and tests, residuals, R-square, estimation/prediction. . . ,
- analysis now completely by **statistical software**, though some explicit formulas exist.

<sup>1</sup> The textbook chapters are very brief — some more details in lecture, full discussion in VHM 8020 and 8120 courses.

<sup>2</sup> PLS 3e Supplement: Sections 26.4-6 (in course syllabus!); S: not covered; IPS 7e: Chapter 13.

<sup>3</sup> Some topics skipped compared to earlier years: Friedman's non-parametric test, ANOVA without replication.

<sup>4</sup> PLS 3e Supplement: Sections 28.1-7 (in part in course syllabus!); S: 10.4; IPS 7e: Chapter 11.

## DATA EXAMPLES FOR TWO-WAY ANOVA AND MULTIPLE REGRESSION

### Energy expenditures in Burkina Faso:<sup>5</sup>

- mean energy expenditures in farming families, divided by gender and season,
- summarized data (no raw data available):

Energy expenditure (calories per day)		Gender	
		men	women
Season	dry	2310	2320
	wet	3460	2890

### Phosphorus levels in tomato plants: (PSLS 3e, Example 26.12)

- 3 levels of nitrogen fertilizer (0/28/160 *kg/ha*) applied to two genotypes of tomato plants (wild/mutant, wild being susceptible to Mycorrhizal fungi),
- genotypes  $\sim$  blocks, 6 replicates per treatment  $\times$  block.

### Test scores for 224 computer science major students in one year:<sup>6</sup>

- $Y_i$  = grade point average (GPA) for first three semesters,
- $x_{i1}, x_{i2}, x_{i3}$  = high school grades in math (HSM), science (HSS), English (HSE),
- $x_{i4}, x_{i5}$  = scholastic aptitude test scores, math part (SATM), verbal part (SATV),
- $x_{i6}$  = sex (1 = men, 2 = women).

for  $i^{\text{th}}$  student,  $i = 1, \dots, 224$ .

---

<sup>5</sup> Based on Payne: Nutrition adaptation in man: social adjustments and their nutritional implications, in Blaxter & Waterlow (eds.): *Nutrition Adaptation in Man.*, Libbey, London, 1985.

<sup>6</sup> CSdata included with earlier IPS textbooks; Campbell & McCabe (1984), *Communications of the ACM*, 1108–1113.

## FACTORIAL DESIGNS

**Factor** (categorical, **explanatory** variable, e.g. treatment/control):

- **grouping of observations** into categories/levels, either by symbols (e.g. letters, roman numbers) or numbers,
- usually, it does not matter if factors are coded by numbers or symbols: **use most natural coding**, and if factors are coded numerically, check df to ensure their modelling as a grouping.<sup>7</sup>

**Several factors in the same design?** — Yes!, in good designs it is possible to separate effects of different factors from each other ⇒

- **cheaper** (less experimental units) than in separate experiments,<sup>8</sup>
- possible to study **combined effect** of several factors,
- **increased scope** of the study/experiment,

and **analyzing multi-factorial data by each factor separately**: is generally **wrong** and only gives valid results if at most one factor is of importance.

Two types of **randomization for factorial experiments**:

- completely randomized design,
- (randomized) block design.

---

<sup>7</sup> The software may misunderstand the factor as continuous and estimate a slope.

<sup>8</sup> The advantage arises e.g. if assessment of nitrogen effects can be done on wild and mutant tomato plants combined, in **additive** models introduced later.

## NOTATION FOR TWO-WAY ANOVA

Data layout  
and notation:

observations $X_{ijk}$	column (C) factor $\sim j$					
	1	...	$j$	...	$J$	
row (R) factor $\sim i$	1	$X_{111}, \dots, X_{11n_{11}}$	...	$X_{1j1}, \dots, X_{1jn_{1j}}$	...	$X_{1J1}, \dots, X_{1Jn_{1J}}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$i$	$X_{i11}, \dots, X_{i1n_{i1}}$	...	$X_{ij1}, \dots, X_{ijn_{ij}}$	...	$X_{iJ1}, \dots, X_{iJn_{iJ}}$	
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$I$	$X_{I11}, \dots, X_{I1n_{I1}}$	...	$X_{Ij1}, \dots, X_{Ijn_{Ij}}$	...	$X_{IJ1}, \dots, X_{IJn_{IJ}}$	

- $X_{ijk}$  =  $k$ th observation in group defined by row factor  $R = i$  and column factor  $C = j$ ,<sup>9</sup>
  - \*  $i = 1, \dots, I$ , and  $I$  = number of levels of R/rows,
  - \*  $i = j, \dots, J$ , and  $J$  = number of levels of C/columns,
  - \*  $k = 1, \dots, n_{ij}$ , and  $n_{ij}$  = number of obs. in  $(i, j)$ th group,
- let also  $N = \sum_{ij} n_{ij}$  the **total number of obs.**, and  $\bar{X} = \sum_{ijk} X_{ijk}/N$  the **overall mean**,
- **terminology**: the dataset/design
  - \* is **balanced**, if all groups equally large ( $n_{11} = \dots = n_{IJ}$ , similar to one-way ANOVA), otherwise unbalanced (also here not necessarily a problem),
  - \* is **complete**, if all  $I \cdot J$  groups present, otherwise incomplete (**difficult** design, avoid if possible!),
  - \* has **replication**, if some of the  $n_{ij}$ 's  $> 1$ , otherwise no replication (all  $n_{ij} = 1$ ).

<sup>9</sup> IPS uses the notation: A=row factor, B=column factor.

## ONE-WAY ANOVA FOR TWO-WAY FACTORIAL

In a two-way design **with replication**, if we focus only on the grouping from the row and column factors ( $I \cdot J$  groups) and otherwise forget about row and column factors  $\Rightarrow$  1-way ANOVA for **combined factor** with  $I \cdot J$  levels:

- **Model:**

$$X_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad \text{for } \varepsilon_{ijk}\text{'s i.i.d. and } \sim N(0, \sigma),$$

and where  $\mu_{ij}$ 's are group (population) means,

- **Estimation:**

$$\hat{\mu}_{ij} = \bar{X}_{ij}. \quad (\text{combined group means}),$$

$$\hat{\sigma}^2 = s_p^2 = \sum_{ij} \frac{n_{ij}-1}{N-IJ} s_{ij}^2 = \text{MSE}, \quad \text{and DFE} = N - IJ,$$

where  $s_{ij}$  = sample standard deviation in group  $(i, j)$ ,

- **ANOVA table:**

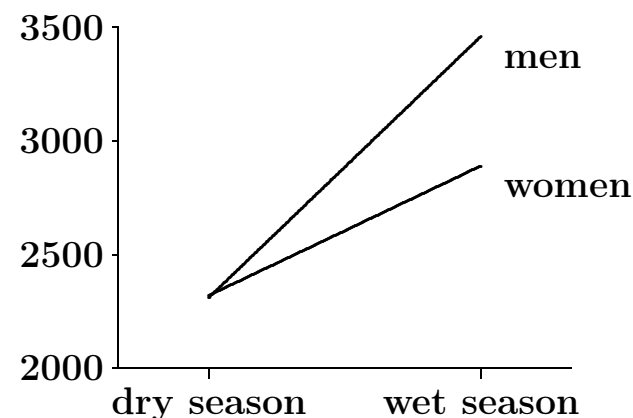
Source	DF	SS	MS	<i>F</i>
Groups	$IJ - 1$	$\sum_{ij} n_{ij} (\bar{X}_{ij} - \bar{X})^2$	SSG/DFG	MSG/MSE
Error	$N - IJ$	$\sum_{ijk} (X_{ijk} - \bar{X}_{ij})^2$	SSE/DFE	
Total	$N - 1$	$\sum_{ijk} (X_{ijk} - \bar{X})^2$		

- **Problem:** analysis does not directly give information about row and column factors separately  $\Rightarrow$  need to **decompose** (split up) model's group terms.

## DECOMPOSING A TWO-WAY TABLE OF MEANS I

**Example:** Energy expenditures in Burkina Faso:

Energy expend. (calories)		Gender		Mean
		men	women	
Season	dry	2310	2320	2315
	wet	3460	2890	3175
Mean		2885	2605	2745



Different ways to look at the data:

- (i) four separate groups,
- (ii) two gender groups for each season, or (iii) two season groups for each gender,
- (iv) (overall level), two season groups, two gender groups, association between gender and season.

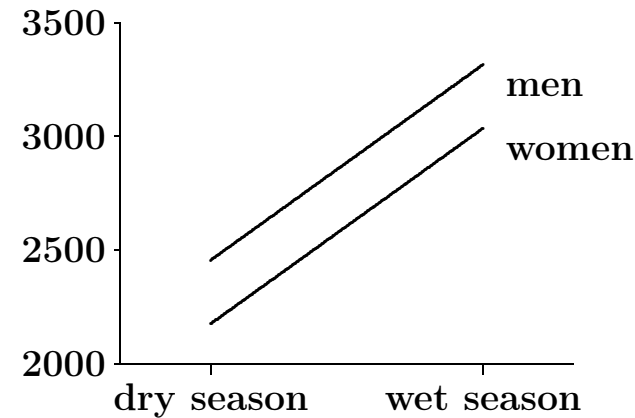
Decomposition of means  $\sim$  (iv):

$\bar{X}$	2745   2745	140   -140	$\bar{X}_{.j} - \bar{X}$
overall mean	2745   2745	140   -140	gender effect
$\bar{X}_{i.} - \bar{X}$	-430   -430	-145   145	$\bar{X}_{ij} - \bar{X}_{i.}$
season effect	430   430	145   -145	$-\bar{X}_{.j} + \bar{X}$

## DECOMPOSING A TWO-WAY TABLE OF MEANS II

Modified energy expenditures  
in Burkina Faso:

Energy expend. (calories)		Gender		Mean
		men	women	
Season	dry	2455	2175	2315
	wet	3315	3035	3175
Mean		2885	2605	2745



Decomposition  
of means  $\sim$  (iv):

$\bar{X}$	2745   2745	140   -140	$\bar{X}_{.j} - \bar{X}$
overall mean	2745   2745	140   -140	gender effect
$\bar{X}_i - \bar{X}$	-430   -430	0   0	$\bar{X}_{ij} - \bar{X}_i$
season effect	430   430	0   0	$-\bar{X}_{.j} + \bar{X}$

Comparison of two variants of the data:

- same overall level, same overall (average) effects of gender and season,
- **modified data**: parallel lines  $\Rightarrow$  **additive** effects — same effect of one factor at all levels of other factor(s),
- **original data**: non-parallel lines  $\Rightarrow$  **non-additive** effects, or **interaction** between the factors gender and season.

## INTERACTION AND ADDITIVITY

### Interaction — some other words:

- synergism or antagonism (depending on the type of interaction),
- epistasy (genetics),
- covariation and association.<sup>10</sup>

### Interaction between two factors:

- the main effects provide an incomplete description, i.e.: the **combined effect of two factors** is not predictable from the isolated effect of each of them when examined separately,
- **the effect of the first factor depends on the level of the second factor** (or vice versa) — “it depends...”,
- **no additivity** between factors  $\Rightarrow$  **interaction** and **additivity** are **opposites**,
- **non-parallel lines** ( $\Rightarrow$  parallel lines  $\sim$  additivity).

---

<sup>10</sup> Both of these terms can be misunderstood as referring to the relation between the two factors themselves, instead of to their combined impact on an outcome.

## TWO-WAY ANOVA MODELS

**Basic model** (in two equivalent formulations):

$$X_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

where the random terms (“errors”)  $\varepsilon_{ijk}$  are i.i.d. and  $\sim N(0, \sigma)$ .

**New parameters** and interpretations:<sup>11</sup>

- $\mu$  = overall mean,
- $\alpha_i$  = “main effect” of  $i$ th row group,
- $\beta_j$  = “main effect” of  $j$ th column group,
- $\gamma_{ij}$  = “interaction effect” of  $(i, j)$ th group (only meaningful with replication).

**Overview of models:**<sup>12</sup> (where  $R \sim$  row factor, and  $C \sim$  column factor)

$EX_{ij}$	Model formula	Interpretation	Corresponding model
$\mu_{ij}$	$(\mu+) R+C+R*C$	interaction $R*C$	one-way ANOVA for $R \times C$
$\mu + \alpha_i + \beta_j$	$(\mu+) R+C$	additivity	<b>new</b> model
$\mu + \alpha_i$	$(\mu+) R$	no effect of $C$	one-way ANOVA for $R$
$\mu + \beta_j$	$(\mu+) C$	no effect of $R$	one-way ANOVA for $C$
$\mu$	$(\mu)$	no effects	one-sample analysis

<sup>11</sup> Technical note: it is necessary to put some **restrictions** on  $\alpha$ 's,  $\beta$ 's and  $\gamma$ 's (otherwise too many parameters).

<sup>12</sup> Final models, i.e. after disregarding non-significant terms

## ANALYSIS OF TWO-WAY ANOVA (WITH REPLICATION)

- same steps as in one-way ANOVA: estimation, model check, ANOVA table with  $F$ -tests, contrasts and/or graphical presentation,<sup>13</sup>
- more rows in ANOVA table, because of more tests.

Two-way ANOVA table:

Source of variation	Degrees of freedom	Sum of squares	Mean square	Hypothesis/ $F$ -statistic
Row factor R	$I - 1$	$\sum_{ij} n_{ij} (\bar{X}_{i.} - \bar{X})^2$	SSR/DFR	$H_0$ : no row eff. $F = \text{MSR}/\text{MSE}$
Column factor C	$J - 1$	$\sum_{ij} n_{ij} (\bar{X}_{.j} - \bar{X})^2$	SSC/DFC	$H_0$ : no column eff. $F = \text{MSC}/\text{MSE}$
Interaction R×C	$(I - 1)(J - 1)$	SSG - SSR - SSC	SSRC/DFRC	$H_0$ : no interaction $F = \text{MSRC}/\text{MSE}$
Error	$N - IJ$	$\sum_{ijk} (X_{ijk} - \bar{X}_{ij})^2$	SSE/DFE	
Total	$N - 1$	$\sum_{ijk} (X_{ijk} - \bar{X})^2$		

- test for interaction first (“read ANOVA table from bottom to top”): if significant, base conclusions on  $\hat{\mu}_{ij}$ ’s ( $\bar{X}_{ij}$ .’s) alone (using pairwise comparisons or contrasts),<sup>14</sup>
- additive model estimation and interpretation: separately for row and column factors (based on respective tests, and row and column means).

<sup>13</sup> Also, as in all linear models, the error standard deviation  $\sigma$  is estimated as  $\sqrt{\text{MSE}}$  with DFE degrees of freedom.

<sup>14</sup> In the presence of strong interaction, tests for main effects are often not meaningful!

## SUMMARY OF TWO-WAY ANOVA FOR TOMATO DATA

Statistical model:

$$X_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

where  $i = 1, 2, 3$  (nitrogen: 0, 28, 160),  $j = 1, 2$  (mutant, wild), and  $k = 1, \dots, 6$ , or as a model formula:

$$\ln(\text{Phosphorus}) = \text{N} + \text{Type} + \text{N*Type} + \text{Error}.$$

ANOVA table:

Source	DF	SS	MS	F	P-value
Nitrogen	2	0.9171	0.4586	28.4	<0.0005
Genotype	1	3.9654	3.9654	246	<0.0005
Interaction N*G	2	0.0536	0.0268	1.66	0.21
Error	30	0.4843	0.0161		
Total	35	5.4204	$s = \sqrt{\text{MSE}} = 0.127$		

strong significance for Nitrogen  
strong significance for Genotype  
interaction non-significant

Presentation (on log-scale): (using  $t^* = t_{.975}(30) = 2.042$ )

statistic	Nitrogen			Genotype	
	0	28	160	Mutant	Wild
$\bar{X}_{i..}$ or $\bar{X}_{.j}$	-1.040	-1.223	-1.431	-1.563	-0.900
95% CI	$\pm t^* s / \sqrt{12} = \pm 0.075$			$\pm t^* s / \sqrt{18} = \pm 0.061$	
LSD <sub>0.95</sub>	$t^* s \sqrt{2/12} = 0.106$			N/A	

**Conclusions:** no evidence of interaction (on log-scale!),  
clear difference between genotypes (Wild > Mutant),  
clear differences between nitrogen levels: decreasing with nitrogen.

## EXERCISES 13.3 AND 13.4

**Exercise 13.3:** (response, factors, number of replications,  $I$ ,  $J$ ,  $N$ )

- (a) **response** = number of hours of sleep “on a typical night”,  
**factors**: smoking categories ( $I = 3$ ), gender ( $J = 2$ ),  
 $n_{ij} = 120$ , and  $N = 720$ ,
- (b) **response** = strength of concrete specimens,  
**factors**: mixtures ( $I = 4$ ), cycles of freezing and thawing ( $J = 3$ ),  
 $n_{ij} = 2$ , and  $N = 24$ ,
- (c) **response** = score on final exam,  
**factors**: teaching methods ( $I = 3$ ), student’s subject of study ( $J = 2$ ),  
 $n_{ij} = 7$  and  $N = 42$ .

**Exercise 13.4:** (sources and degrees of freedom)

- (a) smoking categories (DF = 2), gender (DF = 1), interaction (DF = 2), error (DFE = 714) and total (DFT = 719),
- (b) mixtures (DF = 3), cycles (DF = 2), interaction (DF = 6), error (DFE = 12) and total (DFT = 23),
- (c) teaching methods (DF = 2), study subject (DF = 1), interaction (DF = 2), error (DFE = 36) and total (DFT = 41).

## MULTIPLE LINEAR REGRESSION MODEL

**Data** (such as CSdata) with one outcome  $Y$  and multiple predictors  $x_1, x_2, \dots$ ,

- $x_j$  can be **quantitative** or **categorical**, but here we discuss only quantitative predictors,
- $x_j$  can be **explanatory** or **response**, but response variables are considered as fixed for the modelling.

**Purpose:** use  $x$ -variables to predict the outcome (say GPA), hoping that prediction will be “valid” (i.e., meaningful) for a wider population (of students).

**Alternative purpose:** examine “effect” of  $x$ -variables on GPA (sign, strength, significance of effects), but recall that inferring causal effects from observational data is always **difficult**...

**Multiple linear regression model**<sup>15</sup> (with two predictors for a start):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad \text{or}$$
$$\text{GPA}_i = \beta_0 + \beta_1 \text{HSM}_i + \beta_2 \text{HSS}_i + \varepsilon_i,$$

where the errors  $\varepsilon_1, \dots, \varepsilon_{224}$  are i.i.d. and  $\sim N(0, \sigma^2)$ ,

- same setup as for simple linear regression, but more predictors.

---

<sup>15</sup> Linear, because parameters are only added and multiplied by constants, not necessarily because all predictors enter in the form  $\beta \times x$ .

## MODEL ASSUMPTIONS AND INTERPRETATIONS

Model assumptions:

- independence, normality, variance homogeneity of  $\varepsilon_i$ 's,

- linear relation:

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},$$

$$\hat{Y} = 0.740 + 0.176 \text{HSM} + 0.054 \text{HSS},$$

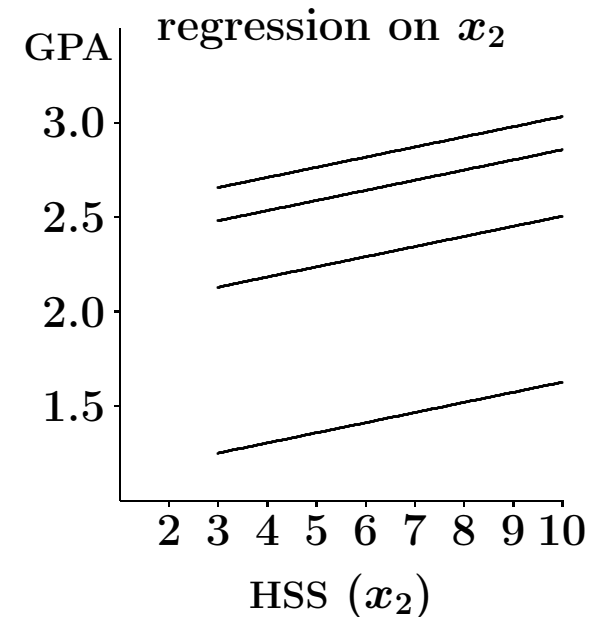
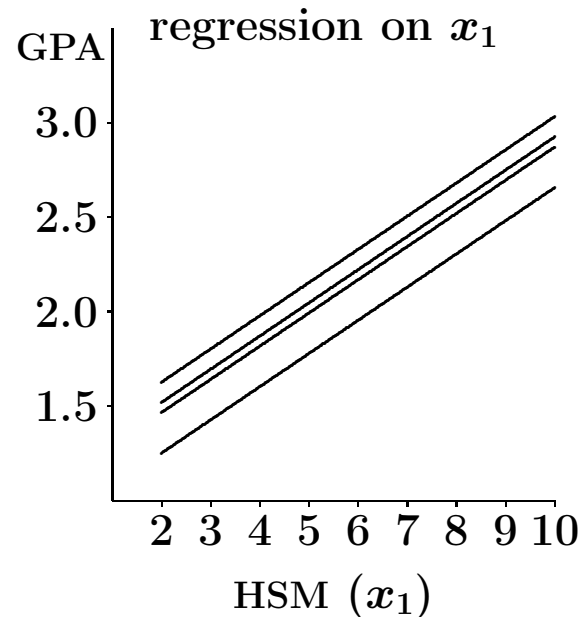
- \* intercept  $\beta_0$  (0.740)  $\sim$  value for  $x_1=0$  and  $x_2=0$ ,

- \* assumes a linear “effect” of  $x_1$  on  $Y$  (for fixed  $x_2$ ),

- \* assumes a linear “effect” of  $x_2$  on  $Y$  (for fixed  $x_1$ ),

- \* assumes additive “effects” of  $x_1$  and  $x_2$  (parallel lines  $\rightarrow$  graph).

Fitted graphs of separate regressions (with other variable fixed at values: min, Q1, median, and Q3/max):



## MULTIPLE LINEAR REGRESSION ANALYSIS

Methods almost the same (as in simple linear regression):

- **estimation** by least squares method (minimizing squared deviations between observed and predicted values),
- **test of simple hypothesis**  $H_0: \beta_j = b$  (for  $b$  known) in the “usual way” (using estimate, SE, reference  $t$ -distribution (DFE)),
- $DFE = n - p - 1$ , where  $n = \#$  observations,  $p = \#$   $x$ -variables in model,
- **confidence and prediction intervals** in the “usual way” (using estimate, SE,  $t^*$ ; for predictions also new values of all  $x$ 's),
- **analysis of variance** (ANOVA) table:
  - \*  $F$ -test is for hypothesis  $H_0: \text{all } \beta_j = 0$  (except  $\beta_0$ ), against alternative  $H_a: \text{some } \beta_j \neq 0$  (not necessarily all) — now different than  $t$ -tests for individual  $\beta$ 's,
  - \* row for Regression (Model) gives variation (SSM) and degrees of freedom (DFM) accounted for by model,
  - \*  $R^2 = SSM/SST \sim$  proportion of variance explained by model, or squared correlation between observed and **fitted** values,
  - \*  $s^2 = \text{MSE}$  and DFE still in row for Residual (Error),
- **model checking** using the residuals: plot now also against each  $x$ -variable to assess linearity.

## INTERPRETATION OF REGRESSION COEFFICIENTS

Each regression coefficient (i.e., the  $\beta_j$  for  $x_j$ ) must be viewed and **interpreted in the presence of other predictors** — and usually changes if the model changes!

- **illustration** by data example:
  - \*  $\hat{\beta}_2 = 0.054$  (.034) for HSS in model with HSM,
  - \*  $\hat{\beta}_1 = 0.151$  (.029) for HSS in model without HSM, i.e. in simple linear regression (sometimes called the crude or univariate estimate for HSS), $\Rightarrow$  size (possibly also the sign) of effects, SE's and significance may change,
- **proper interpretation** for  $\beta_2$ :
  - \*  $\sim$  “effect” of  $x_2$ , when  $x_1$  has been accounted for,
  - \*  $\sim$  additional “effect” of adding  $x_2$  to model with  $x_1$ .

But **when** and **why** does the “effect” of one predictor depend on the others (being included or not)?

- it happens (strongly)<sup>16</sup> when the  $x$ -variables are (strongly) associated/dependent; this we can assess by their correlations, e.g. in data example:  $\text{Corr}(x_1, x_2) = 0.58$ ,
- it happens because dependent  $x$ 's explain some of the **same variation in  $Y$** ,
- **not the same** as interaction, because the model is still assumed additive.

<sup>16</sup> The term **collinearity** is used for a situation with strong dependence among  $x$ -variables; not discussed further here.

## MODEL BUILDING / VARIABLE SELECTION

Given an outcome  $Y$  and a set of predictors  $x_1, \dots, x_p$ : **how to find a good model?**

— we usually only want to include the “important” predictors. . . .

- not necessarily easy to find “best” or even a good model,
- can rarely be done automatically (by some rules or procedures),
- basically, an **exploratory process**,
- interest is in the **most succinct and parsimonious** model,<sup>17</sup>
- but some **guidelines** exist:
  - \* start with simple associations between  $Y$  and each of the predictors (shows which of the predictors are important **on their own**),
  - \* compute also correlations among the  $x$ 's (to see if any of them are strongly correlated, thereby likely to cause trouble),
  - \* when no. of predictors  $\ll$  no. of observations: analysis should start with “full model” including all (important) predictors (and possibly also interactions),
  - \* from a satisfactory “full model”, remove in a stepwise manner non-significant predictors to achieve a final model with only significant predictors.
  - \* **model checking** should be based on the full model (possibly repeated for final model).

---

<sup>17</sup>One might think this to be the model with highest  $R^2$ , however adding variables always  $\Rightarrow$  more variation explained and higher  $R^2$ .

## SUMMARY OF ANALYSIS FOR CSDATA EXAMPLE

- **correlations with GPA:** range 0.44 – 0.11 (HSM – SATV), all significant except last one ( $P = 0.09$ ),
- **correlations between  $x$ 's:** moderate correlations (0.45 – 0.58) among HS-variables and among SAT-variables, less between HS-variables and SAT-variables,
- **simple  $t$ -test** for comparison of GPA between sexes: non-significant,
- **full model:**
  - \* HSM strongly significant, all others non-significant,
  - \* 14 standardized residuals outside  $(-2, 2)$  (11 expected),
  - \* no single, extreme residuals (most extreme:  $-3.05$ ),
  - \* residual plots look satisfactory (not great),

○ **table of model reductions:**

- **final model** with predictors HSM and HSE; estimated equation:

$$\widehat{\text{GPA}} = 0.624 + 0.183 \cdot \text{HSM} + 0.0607 \cdot \text{HSE}.$$

Model	$F$	$P$	$R^2$	$s$	reduction	
					$t$	$P$
full: $x_1 - x_6$	9.72	<0.001	21.2%	0.701	—	—
$x_1 - x_5$	11.7	<0.001	21.1%	0.700	0.29	0.77
$x_1 - x_4$	14.5	<0.001	21.0%	0.699	-0.69	0.49
$x_1, x_3, x_4$	19.1	<0.001	20.7%	0.699	0.88	0.38
$x_1, x_3$	27.9	<0.001	20.2%	0.700	1.22	0.22
$x_1$	52.3	<0.001	19.1%	0.703	1.75	0.08

## MODEL CHECKING FOR ANOVA AND MULTIPLE REGRESSION

### Model assumptions:

same as for previous models (independence, normality, variance homogeneity, mean relation  $\sim$  linearity, equal means within groups, additivity...).

### ANOVA models with multiple factors and multiple regression models:

- often **few replications**  $\Rightarrow$  difficult to check model assumptions separately for each group,
- use instead residuals (“**observed – expected**”) to check model assumptions, similar to previous models:
  - \* **variance homogeneity:**  
plot standardized residuals against model’s fitted/expected values and look for unequal variances across range of fitted values,<sup>18</sup>
  - \* **normal distribution:**  
normal probability plot of standardized residuals,
  - \* **outliers:**  
look for extreme standardized residuals (with same rules as previously),
  - \* **other model violations:**  
plot standardized residuals against data order (if applicable) or other variables.

---

<sup>18</sup> In a multifactorial ANOVA with replication, the one-way ANOVA tools: i) max/min rule, ii) variance test, still apply, when groups are defined by combinations of all factors.

## OVERVIEW ONE-WAY & TWO-WAY ANOVA

### o Data description:

- \* statistics: group<sup>19</sup> means and standard deviations,
- \* graphs: box-plot for groups<sup>19</sup>, interaction plot (two-way),

### o Statistical model:

$$\begin{aligned}\text{one-way : } X_{ij} &= \mu + \alpha_i + \varepsilon_{ij}, \\ \text{two-way (replication) : } X_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \\ \text{two-way (no replication) : } X_{ij} &= \mu + \alpha_i + \beta_j + \varepsilon_{ij},\end{aligned}$$

### o Model checks — residuals plots and/or

- \* equal variance: i) max/min rule, ii) variance test,
- \* normality: normal plots/tests,<sup>20</sup>

### o Statistical analysis:

- \* estimation of pooled (or error) standard deviation,
- \* hypothesis testing of overall effects → ANOVA table,
- \* (two-way with replication): interaction significant/“substantial”?:
  - **yes**: interaction plot, one-way ANOVA for groups<sup>19</sup>,
  - **no**: row and column factors assessed separately,
- \* pairwise comparisons between group means for significant effects (based on CI, LSD or *t*-tests): unadjusted or adjusted to simultaneous error level of 0.05,

### o Presentation: group means with SE or CI + significance indications.

---

<sup>19</sup> In a two-way design with replication, groups refer to the combined groups formed by row and column factors.

<sup>20</sup> With ample replication: observations within groups<sup>19</sup>; With limited/no replication: (standardized) residuals across all groups.

## SUMMARY NOTES

Key words and concepts for two-way ANOVA and multiple regression:

- multifactorial designs:
  - \* **advantages** over single factor designs: larger scope, allows assessment of combined effects, potentially more efficient,
  - \* **characterizations**: factors, factorial notation (e.g.  $2 \times 2$ -design), balancedness, completeness, replication,
  - \* **interaction**: non-parallel curves, non-additive effects, effect of one factor depends on another factor; **opposite of additivity**,
- multiple regression:
  - \* **interpretation** (and values) of regression coefficients depend on other predictors in model, association between predictors,
  - \* **additivity** assumption for multiple predictors (most commonly),
  - \* model building/variable selection,
- **analysis**: ANOVA table,  $F$ -tests for different hypotheses: testing interaction first, interaction plot, model checking by **residuals**, post-ANOVA analysis using LSD-values and pairwise comparisons (possibly with Bonferroni adjustments),  $t$ -tests for individual regression coefficients.