Allowing for the structure of a designed experiment when estimating and testing trait correlations

Hans-Peter Piepho

Biostatistics Unit
Institute of Crop Science
University of Hohenheim

Table of contents

1. Introduction

- ANOVA for designed experiments is based in a linear model
- Pairwise (bivariate) correlations often computed without a model
  $\Rightarrow$ this ignores effects that are accounted for in univariate analyses
- This paper shows how univariate and bivariate analyses can be reconciled

1. Introduction

Example 1

- Thirteen different rootstocks
- 8 apple trees per rootstock
- Completely randomized design
- Traits:
  - trunk girth at 4 years (mm × 100)
  - weight of tree above ground at 15 years (lb × 1000)
- Univariate analysis: one-way ANOVA

(Andrew & Herzberg 1985, p.357-360)

Sample correlation:

$$r = 0.7490$$

1. Introduction

Fig. 1. Scatter plot of trunk girth at 4 years (mm × 100) versus tree weight at 15 years (lb × 1000) for apple (*Malus domestica* L.) rootstock data from an experiment with thirteen rootstocks (Example 1).

2. Linear modelling

$$y_{ij} = \mu + \tau_i + e_{ij} \,, \qquad (1)$$

where

y_{ij} = response of the *j*-th replicate of the *i*-th treatment

μ = general intercept

τ_i = effect of the *i*-th treatment

e_{ij} = residual error term associated with the response y_{ij}

$$e_{ij} \sim N\left(0, \sigma_e^2\right) \qquad (2)$$

2. Linear modelling

Bivariate model for two traits

$$y_{ij1} = \mu_1 + \tau_{i1} + e_{ij1} \text{ and} \qquad (4)$$

$$y_{ij2} = \mu_2 + \tau_{i2} + e_{ij2} \qquad (5)$$

$$\mathrm{var}\left(e_{ij1}\right) = \sigma^2_{e(1)} \qquad (6)$$

$$\mathrm{var}\left(e_{ij2}\right) = \sigma^2_{e(2)} \qquad (7)$$

$$\mathrm{cov}\left(e_{ij1}, e_{ij2}\right) = \sigma_{e(1,2)} \qquad (8)$$

2. Linear modelling

$$\begin{pmatrix} y_{ij1} \\ y_{ij2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \tau_{i1} \\ \tau_{i2} \end{pmatrix} + \begin{pmatrix} e_{ij1} \\ e_{ij2} \end{pmatrix} \qquad (10)$$

$$\begin{pmatrix} e_{ij1} \\ e_{ij2} \end{pmatrix} \sim BVN\left[\phi = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma^2_{e(1)} & \sigma_{e(1,2)} \\ \sigma_{e(1,2)} & \sigma^2_{e(2)} \end{pmatrix}\right] \qquad (11)$$

$$\begin{pmatrix} y_{ij1} \\ y_{ij2} \end{pmatrix} \sim BVN\left[\begin{pmatrix} \mu_1 + \tau_{i1} \\ \mu_2 + \tau_{i2} \end{pmatrix}, \begin{pmatrix} \sigma^2_{e(1)} & \sigma_{e(1,2)} \\ \sigma_{e(1,2)} & \sigma^2_{e(2)} \end{pmatrix}\right] \qquad (12)$$

2. Linear modelling

The bivariate model (12) is consistent with the univariate model (3)

$\Rightarrow$ both comprise the same set of parameters (effects and variances)

Question:

What is the implicit assumption for the distribution of the bivariate observations (y_{ij1}, y_{ij2}), if we compute simple sample correlations *r* from the observed data and subject them to a significance test?

2. Linear modelling

Model assumed for the usual correlation test

$$\begin{pmatrix} y_{ij1} \\ y_{ij2} \end{pmatrix} \sim BVN\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma^2_{e(1)} & \sigma_{e(1,2)} \\ \sigma_{e(1,2)} & \sigma^2_{e(2)} \end{pmatrix}\right] \qquad (13)$$

Test statistic

$t = r\sqrt{n-2}\Big/\sqrt{1-r^2}$ where *n* is the sample size

or

the equivalent F-statistic $F = t^2$

3. Variance and covariance of treatment effects

"Sample variance" of treatment effects:

$$Q(\tau) = \frac{\sum_{i=1}^{t} \left(\tau_i - \bar{\tau}_{\bullet}\right)^2}{t-1} \qquad (14)$$

Natural measure of variability of the fixed treatment effects

(Winer *et al*. 1991, p. 85; Gelman 2005)

3. Variance and covariance of treatment effects

$Q(\tau)$ can be estimated from the two ANOVA mean squares as

$$\hat{Q}(\tau) = \frac{MS_{treat} - MS_{error}}{r} \qquad (15)$$

where

$$MS_{treat} = SS_{treat} / (t-1)$$

$$MS_{error} = SS_{error} / [t(r-1)] = \hat{\sigma}_e^2$$

3. Variance and covariance of treatment effects

Table 1: One-way ANOVA with expected SS; fixed treatment effects.

Source	D.F.	Sum of squares (SS)	E(SS), τ_i fixed §
Treatments	$t-1$	$SS_{Treat} = r\sum_{i=1}^{t}\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2$	$(t-1)\left[\sigma_e^2 + rQ(\tau)\right]$
Error	$t(r-1)$	$SS_{Error} = \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{i\bullet}\right)^2$	$t(r-1)\sigma_e^2$
Corrected total	$rt-1$	$SS_{Total} = SS_{Treat} + SS_{Error}$ $= \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2$	$(rt-1)\sigma_e^2 + (t-1)rQ(\tau)$

& *t* = number of treatments; *r* = number of replications per treatment

§ $Q(\tau) = \sum_{i=1}^{t}\left(\tau_i - \bar{\tau}_{\bullet}\right)^2 / (t-1)$; see, e.g., Winer *et al.* (1991, p. 85)

3. Variance and covariance of treatment effects

Table 2: One-way ANOVA with expected SS; random treatment effects.

Source	D.F.	Sum of squares (SS)	E(SS), τ_i random $
Treatments	$t-1$	$SS_{Treat} = r\sum_{i=1}^{t}\left(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}\right)^2$	$(t-1)\left(\sigma_e^2 + r\sigma_\tau^2\right)$
Error	$t(r-1)$	$SS_{Error} = \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{i\bullet}\right)^2$	$t(r-1)\sigma_e^2$
Corrected total	$rt-1$	$SS_{Total} = SS_{Treat} + SS_{Error}$ $= \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{\bullet\bullet}\right)^2$	$(rt-1)\sigma_e^2 + (t-1)r\sigma_\tau^2$

& t = number of treatments; r = number of replications per treatment

$ See, e.g., Winer *et al*. (1991, p. 92) or Searle *et al*. (1992, p. 60)

3. Variance and covariance of treatment effects

Estimating the treatment variance when treatments are random

$$\hat{\sigma}_{\tau}^{2} = \frac{MS_{treat} - MS_{error}}{r} . \qquad (16)$$

- ANOVA or methods-of-moments estimator of the treatment variance σ_{τ}^{2} (Winer *et al*. 1991, p. 92; Searle *et al*. 1992, p. 59)
- Assuming normality of the treatment effects and non-negativity of the variance estimator, (16) also coincides with the REML estimator of σ_{τ}^{2} (Searle *et al*. 1992, p. 92)

3. Variance and covariance of treatment effects

Most important point:

$$\hat{Q}(\tau) = \hat{\sigma}_{\tau}^{2}$$

3. Variance and covariance of treatment effects

"Sample covariance" of treatment effects

$$Q(\tau_1, \tau_2) = \frac{\sum_{i=1}^{t} (\tau_{i1} - \bar{\tau}_{\bullet 1})(\tau_{i2} - \bar{\tau}_{\bullet 2})}{t-1} \qquad (17)$$

3. Variance and covariance of treatment effects

Estimation:

$$SP_{treat} = r\sum_{i=1}^{t}\left(\bar{y}_{i\bullet 1} - \bar{y}_{\bullet\bullet 1}\right)\left(\bar{y}_{i\bullet 2} - \bar{y}_{\bullet\bullet 2}\right) \qquad (18)$$

$$SP_{error} = \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij1} - \bar{y}_{i\bullet 1}\right)\left(y_{ij2} - \bar{y}_{i\bullet 2}\right) \qquad (19)$$

3. Variance and covariance of treatment effects

The method-of-moments estimator:

$$\hat{Q}(\tau_1, \tau_2) = \frac{MP_{treat} - MP_{error}}{r} \qquad (20)$$

$$\hat{\sigma}_{e(1,2)} = MP_{error} \qquad (21)$$

where

$$MP_{treat} = SP_{treat} / (t-1)$$

$$MP_{error} = SP_{error} / [t(r-1)]$$

3. Variance and covariance of treatment effects

When treatments are taken as random:

$$\sigma_{\tau(1,2)} = \operatorname{cov}(\tau_{i1}, \tau_{i2})$$

$$\hat{\sigma}_{\tau(1,2)} = \hat{Q}(\hat{\tau}_1, \hat{\tau}_2) \Rightarrow \text{REML}$$

$$\hat{\sigma}_{e(1,2)} = MP_{error}$$

3. Variance and covariance of treatment effects

In case of convergence problems with bivariate REML:

$$SP_{error} = \left[SS_{error}\left(y_1 + y_2\right) - SS_{error}\left(y_1\right) - SS_{error}\left(y_2\right)\right]/2 \qquad (22)$$

3. Variance and covariance of treatment effects

Bivariate model with random treatment effects

$$\begin{pmatrix} \tau_{i1} \\ \tau_{i2} \end{pmatrix} \sim BVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\tau(1)} & \sigma_{\tau(1,2)} \\ \sigma_{\tau(1,2)} & \sigma^2_{\tau(2)} \end{pmatrix}\right] \qquad (23)$$

$$corr(\tau_{i1}, \tau_{i2}) = \rho_{\tau(1,2)} = \frac{\sigma_{\tau(1,2)}}{\sigma_{\tau(1)}\sigma_{\tau(2)}} \qquad (24)$$

3. Variance and covariance of treatment effects

Marginal (total) correlation for the observed data $\left(y_{ij1}, y_{ij2}\right)$

$$corr\left(y_{ij1}, y_{ij2}\right) = \rho_{y(1,2)} = \frac{\sigma_{y(1,2)}}{\sigma_{y(1)}\sigma_{y(2)}} \qquad (25)$$

$$\sigma_{y(1)}^2 = \mathrm{var}\left(y_{ij1}\right) = \sigma_{\tau(1)}^2 + \sigma_{e(1)}^2 \qquad (26)$$

$$\sigma_{y(2)}^2 = \mathrm{var}\left(y_{ij2}\right) = \sigma_{\tau(2)}^2 + \sigma_{e(2)}^2 \qquad (27)$$

$$\sigma_{y(1,2)} = \mathrm{cov}\left(y_{ij1}, y_{ij2}\right) = \sigma_{\tau(1,2)} + \sigma_{e(1,2)} \qquad (28)$$

4. Example 1 continued

Parameter	Effect	Estimate	S.E.
$\sigma^2_{\tau(1)}$	Treatment	0.01247	.005422
$\sigma^2_{\tau(2)}$	Treatment	0.1586	0.06781
$\rho_{\tau(1,2)}$	Treatment	0.8908	0.06978
$\sigma^2_{e(1)}$	Error	0.006462	0.000958
$\sigma^2_{e(2)}$	Error	0.06026	0.008933
$\rho_{e(1,2)}$	Error	0.4572	0.08292
$\sigma^2_{y(1)}$	Total	0.01893	
$\sigma^2_{y(2)}$	Total	0.2188	
$\sigma_{y(1,2)}$	Total	0.04863	
$\rho_{y(1,2)}$	Total	0.7556	

Table 3: Variance parameter estimates for the apple (*Malus domestica* L.) rootstock data in Table 1 (Example 1).

Trait 1 = trunk girth at 4 years (mm × 100)
Trait 2 = tree weight at 15 years (lb × 1000)

For comparison
$r = 0.7490$

4. Example 1 continued

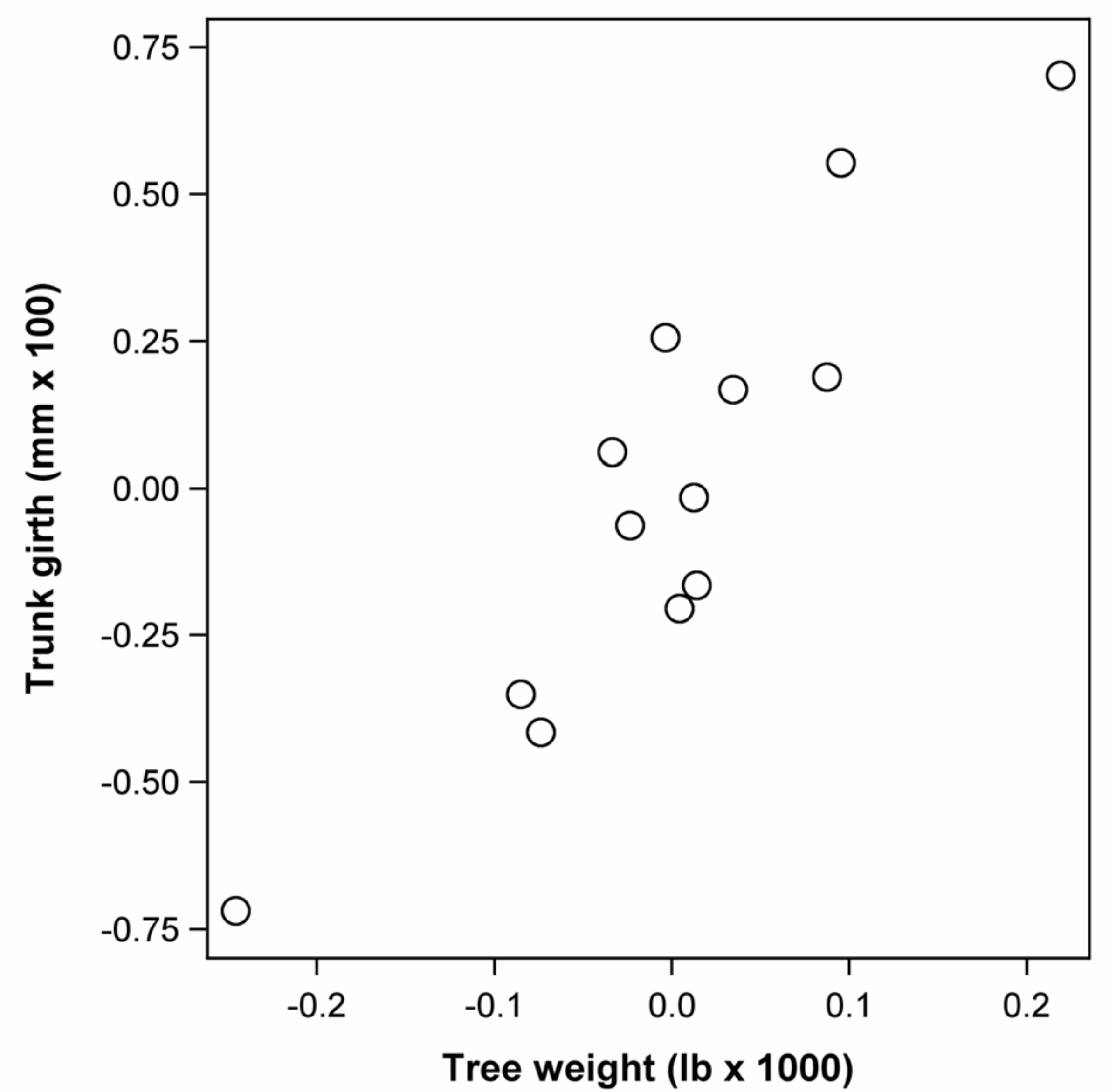


Fig. 2. Scatter plot of BLUPs of treatment effects trunk girth at 4 years (mm × 100) versus tree weight at 15 years (lb × 1000) for apple (*Malus domestica* L.) rootstock data from an experiment with thirteen rootstocks (Example 1).

$$\hat{\rho}_{\tau(1,2)} = 0.8908$$

4. Example 1 continued

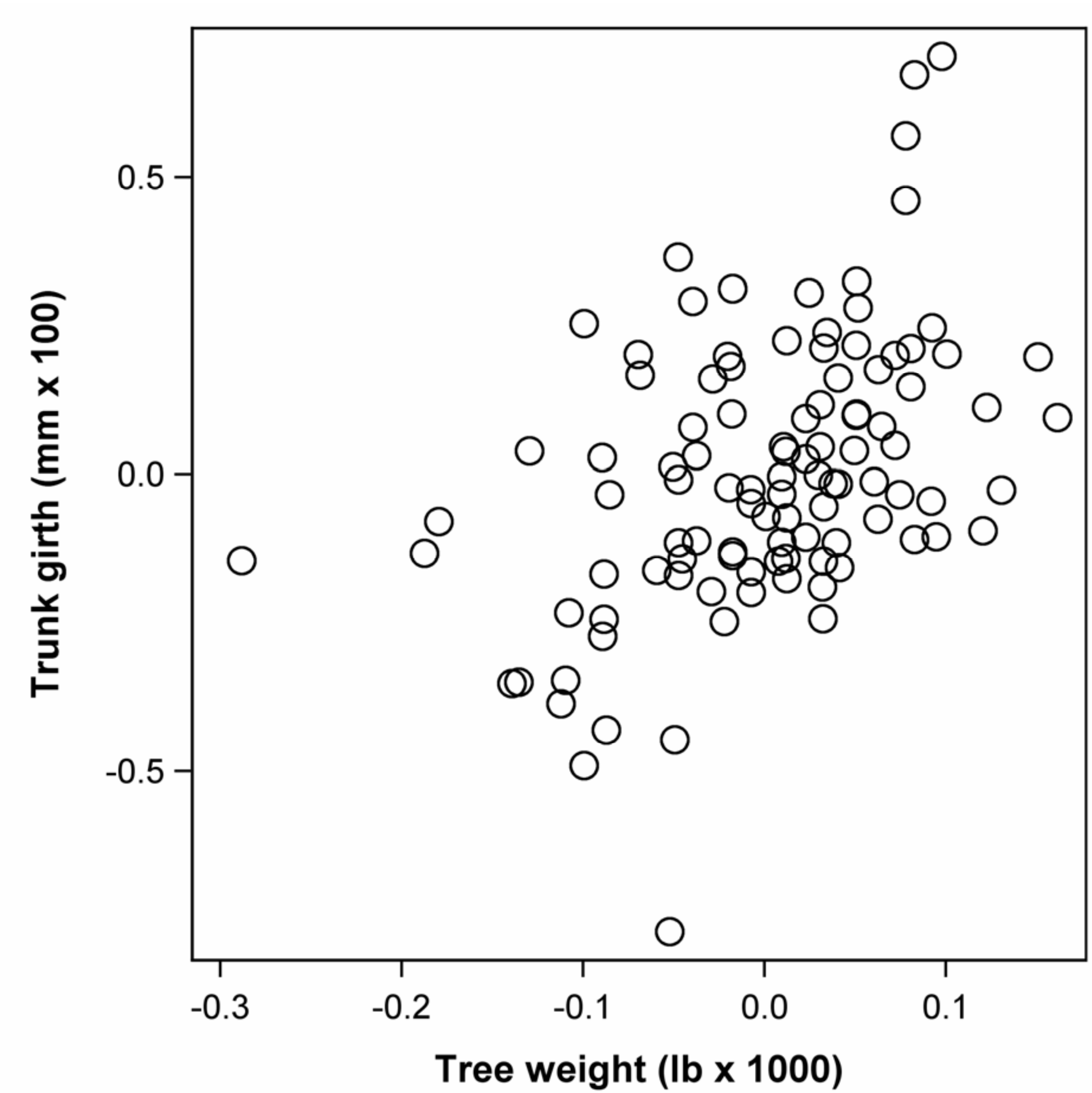


Fig. 3. Scatter plot of residuals for trunk girth at 4 years (mm × 100) versus tree weight at 15 years (lb × 1000) for apple (*Malus domestica* L.) rootstock data from an experiment with thirteen rootstocks (Example 1).

$$\hat{\rho}_{e(1,2)} = 0.4572$$

4. Example 1 continued

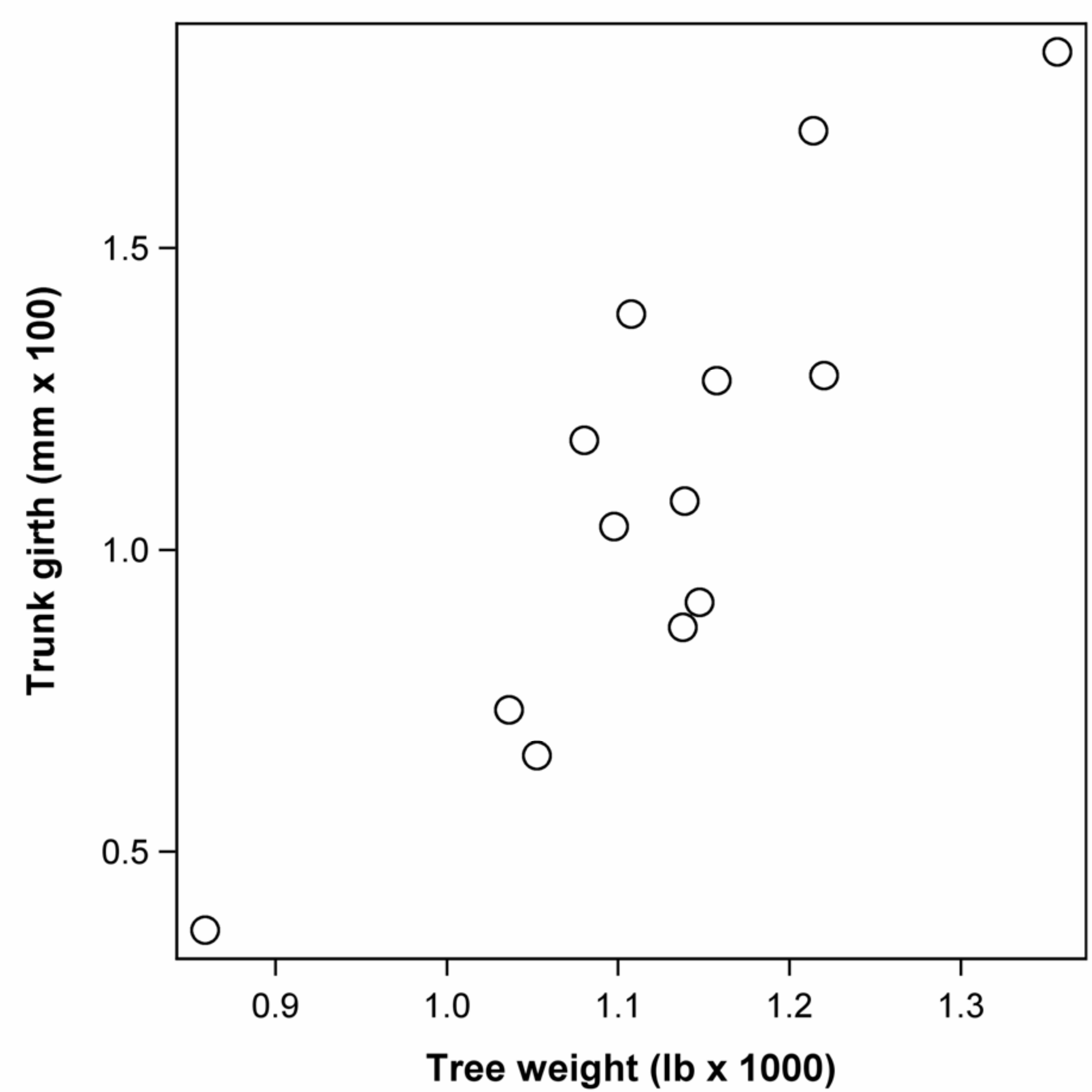


Fig. 4. Scatter plot rootstock means for trunk girth at 4 years (mm × 100) versus tree weight at 15 years (lb × 1000) for apple rootstock data from an experiment with thirteen apple (*Malus domestica* L.) rootstocks (Example 1).

$$r = 0.8675$$

$$\bar{y}_{i\bullet} = \mu + \tau_i + \bar{e}_{i\bullet}$$

4. Example 1 continued

Likelihood ratio tests

$$T = -2\left[\log L(reduced) - \log L(full)\right]$$

Correlation?

$$H_0 : \rho_{e(1,2)} = 0$$

$$T = 21.34\,,\; d.f. = 1\,,\; p < 0.0001$$

4. Example 1 continued

Correlations equal between effects?

$$H_0 : \rho_{e(1,2)} = \rho_{\tau(1,2)}$$

$$T = 7.16\,,\; d.f. = 1\,,\; p = 0.0074$$

5. Example 2

- Field experiment with 64 breeding lines of oats (*Avena sativa* L.)
- Square lattice with three replicates (Friedrich Utz, personal communication).
- Traits:

  - yield (tons/ha $\times$ 10)

  - plant height (cm) (Figure 5).
- Three outliers removed

5. Example 2

Linear mixed model

$$y_{ijk} = \mu + \tau_i + r_k + b_{jk} + e_{ijk}$$

y_{ijk} = response of the *i*-th genotype in the *j*-th incomplete block nested within the *k*-th replicate

r_k = effect of k-th replicate

b_{jk} = effect of *jk*-th block

τ_i = effect of *i*-th treatment

e_{ijk} = error of *ijk*-th plot

5. Example 2

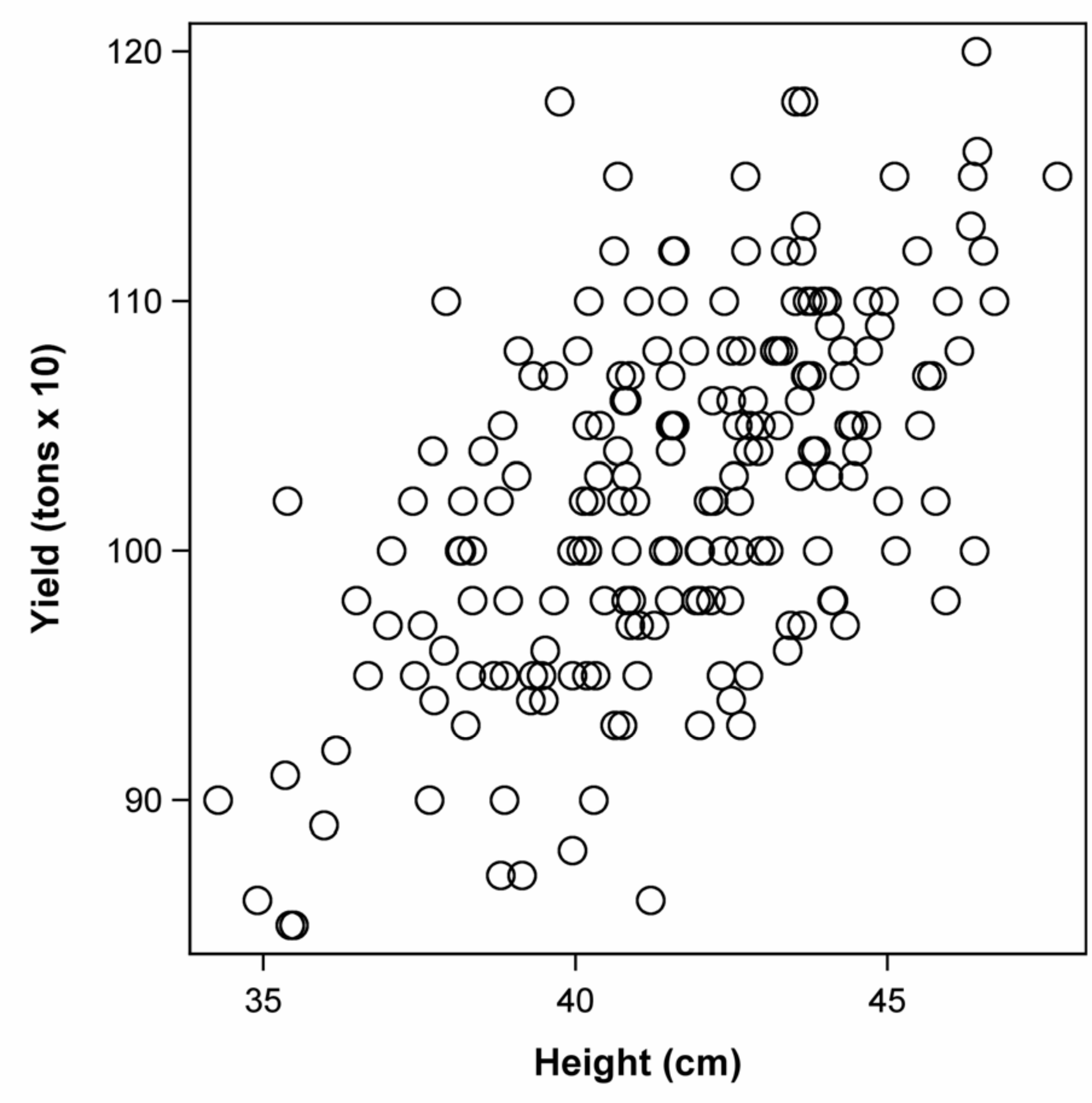


Fig. 5. Scatter plot of yield (kg/ha) versus height (cm) for oats (*Avena sativa* L.) data (Example 2).

$$r = 0.5750$$

5. Example 2

Table 5: Variance parameter estimates for oats (*Avena sativa* L.) data (Example 2). Trait 1 = yield (tons × 10), trait 2 = height (cm).

Parameter	Effect	Estimate	Standard error
$\sigma^2_{\tau(1)}$	Treatment	22.9713	4.7077
$\sigma^2_{\tau(2)}$	Treatment	2.8592	0.6688
$\rho_{\tau(1,2)}$	Treatment	0.4796	0.1207
$\sigma^2_{r(1)}$	Replicate	13.9498	15.6364
$\sigma^2_{r(2)}$	Replicate	1.7517	2.0340
$\rho_{r(1,2)}$	Replicate	0.7661	0.3275

5. Example 2

Parameter	Effect	Estimate	Standard error
$\sigma^2_{b(1)}$	Block	12.3524	4.2842
$\sigma^2_{b(2)}$	Block	1.9970	0.7245
$\rho_{b(1,2)}$	Block	0.9118	0.06440
$\sigma^2_{e(1)}$	Error	8.5396	1.1615
$\sigma^2_{e(2)}$	Error	1.9457	0.2769
$\rho_{e(1,2)}$	Error	0.2795	0.08932
$\sigma^2_{y(1)}$	Total	8.5537	
$\sigma^2_{y(2)}$	Total	57.8131	
$\rho_{y(1,2)}$	Total	0.6000	

5. Example 2

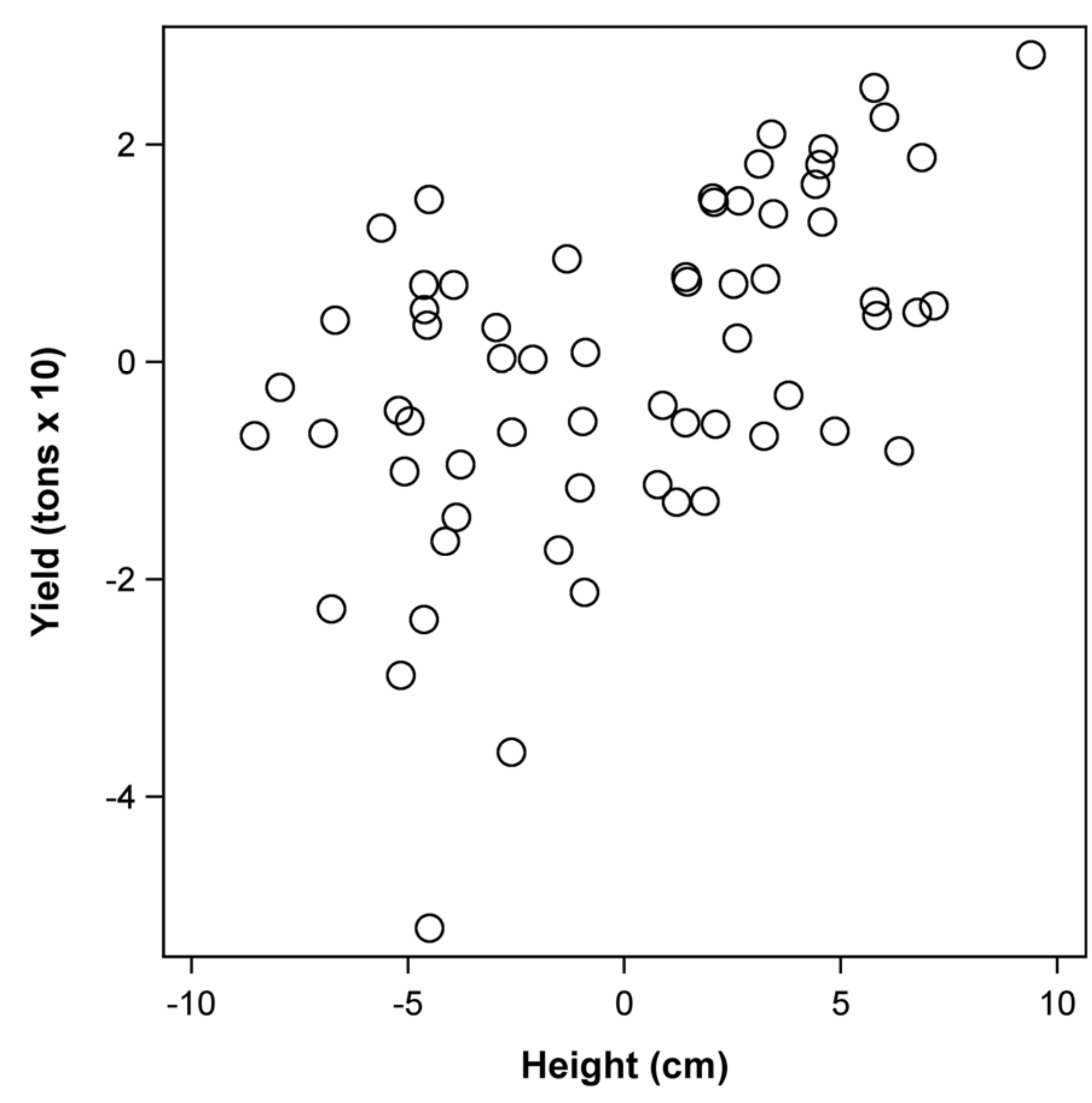


Fig. 6. Scatter plot of BLUPs of treatment effects for yield (kg/ha) versus height (cm) for oats (*Avena sativa* L.) data (Example 2).

$$\hat{\rho}_{\tau(1,2)} = 0.4796$$

5. Example 2

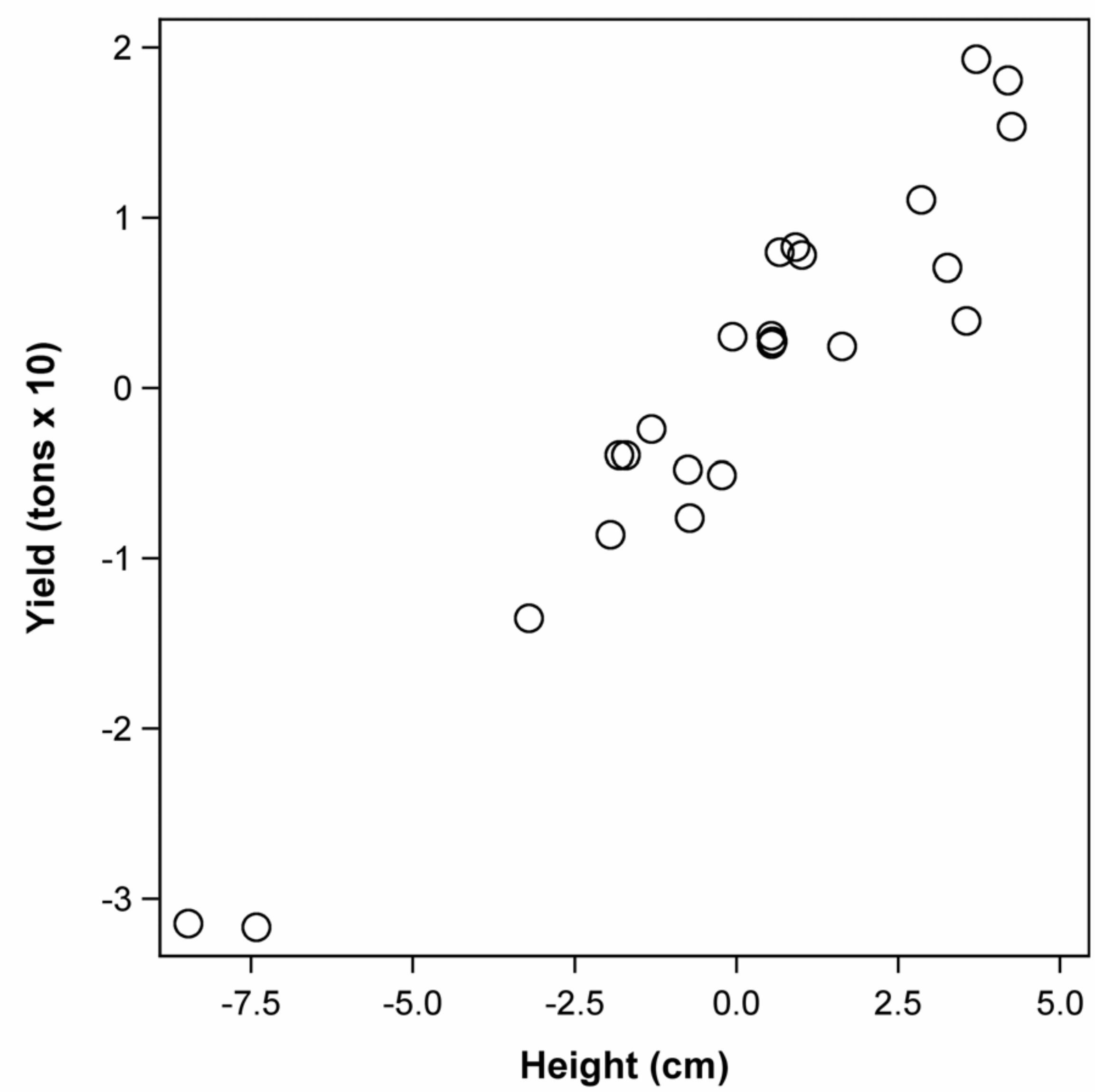


Fig. 7. Scatter plot of BLUPs of block effects for yield (kg/ha) versus height (cm) for oats (*Avena sativa* L.) data (Example 2).

$$\hat{\rho}_{b(1,2)} = 0.9118$$

5. Example 2

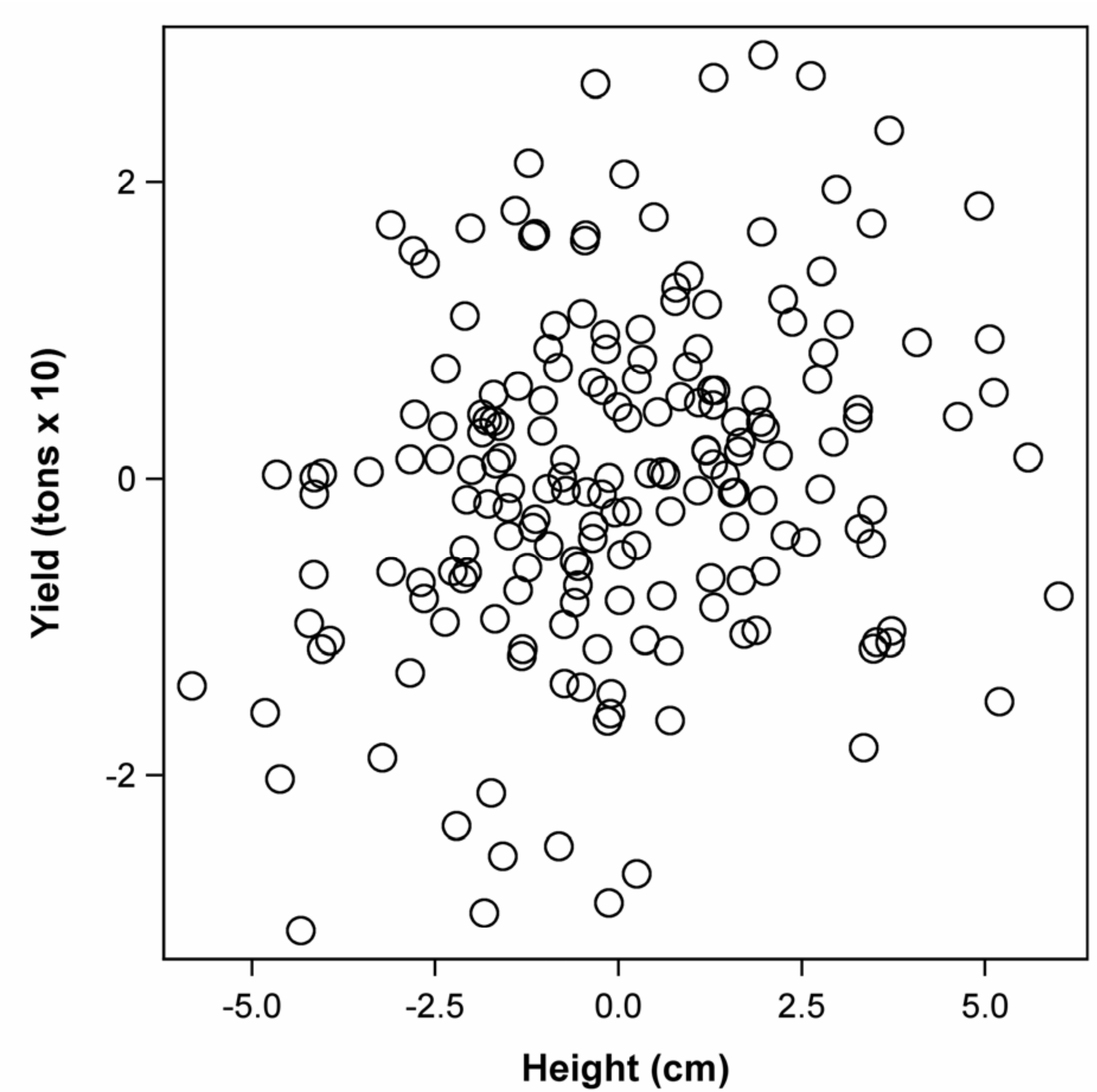
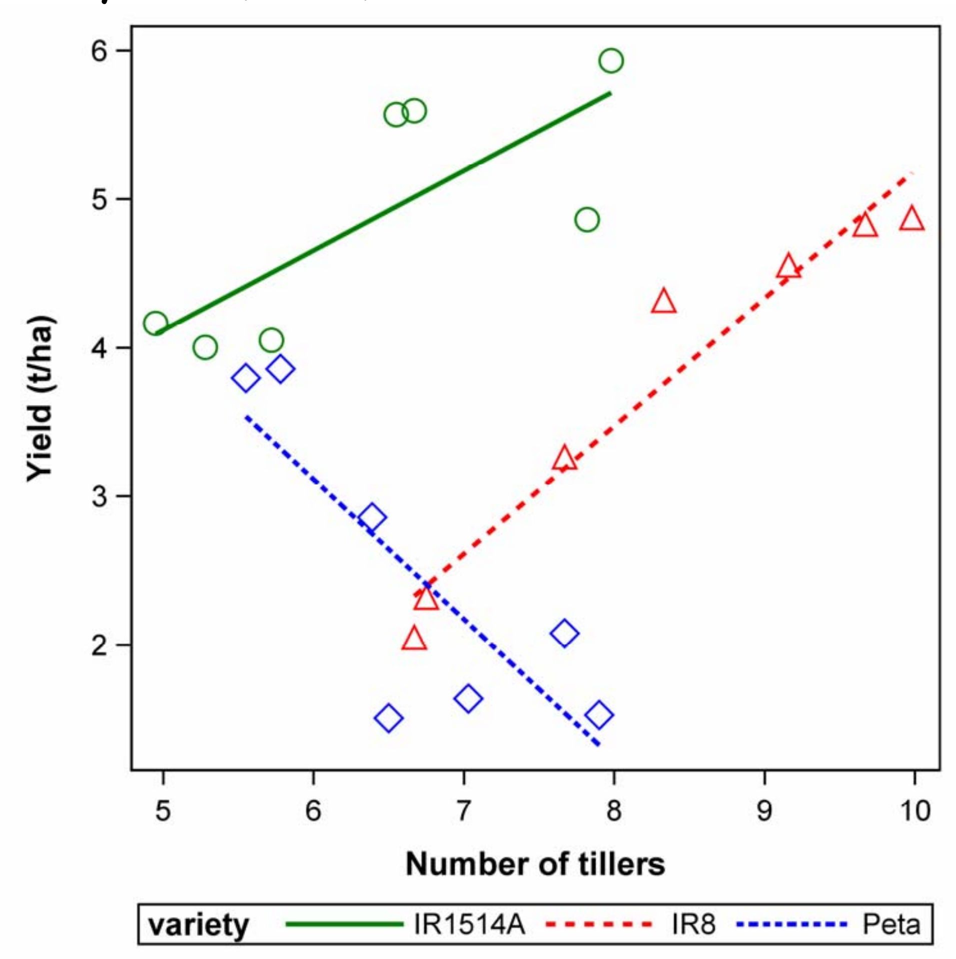


Fig. 8. Scatter plot of residuals for yield (kg/ha) versus height (cm) for oats (*Avena sativa*) data (Example 2).

$$\hat{\rho}_{e(1,2)} = 0.2795$$

6. Example 3

- 3 rice varieties (IR1514A, IR8 und Peta), completely randomized
- yield (t/ha) & tiller number



$$r = 0.1973$$

$$(t = 0.88, p = 0.3912)$$

6. Example 3

Tabelle 6: Residual variance parameter estimates by variety for rice (*Oryza sativa* L.) data of Gomez & Gomez (1984, p. 377) (Example 3) under model with heterogeneity between varieties. Trait 1 = number of tillers, trait 2 = yield of rice (t/ha).

Parameter	Sorte					
	IR1514A		IR8		Peta	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
$\sigma^2_{e(1)}$	1.3688	0.7731	1.8521	1.0964	0.7862	0.4473
$\sigma^2_{e(2)}$	0.6649	0.3772	1.4543	0.8583	1.0692	0.6105
$\rho_{e(1,2)}$	0.7665	0.1667	0.9702	0.0242	-0.8061	0.1421

7. Example 4

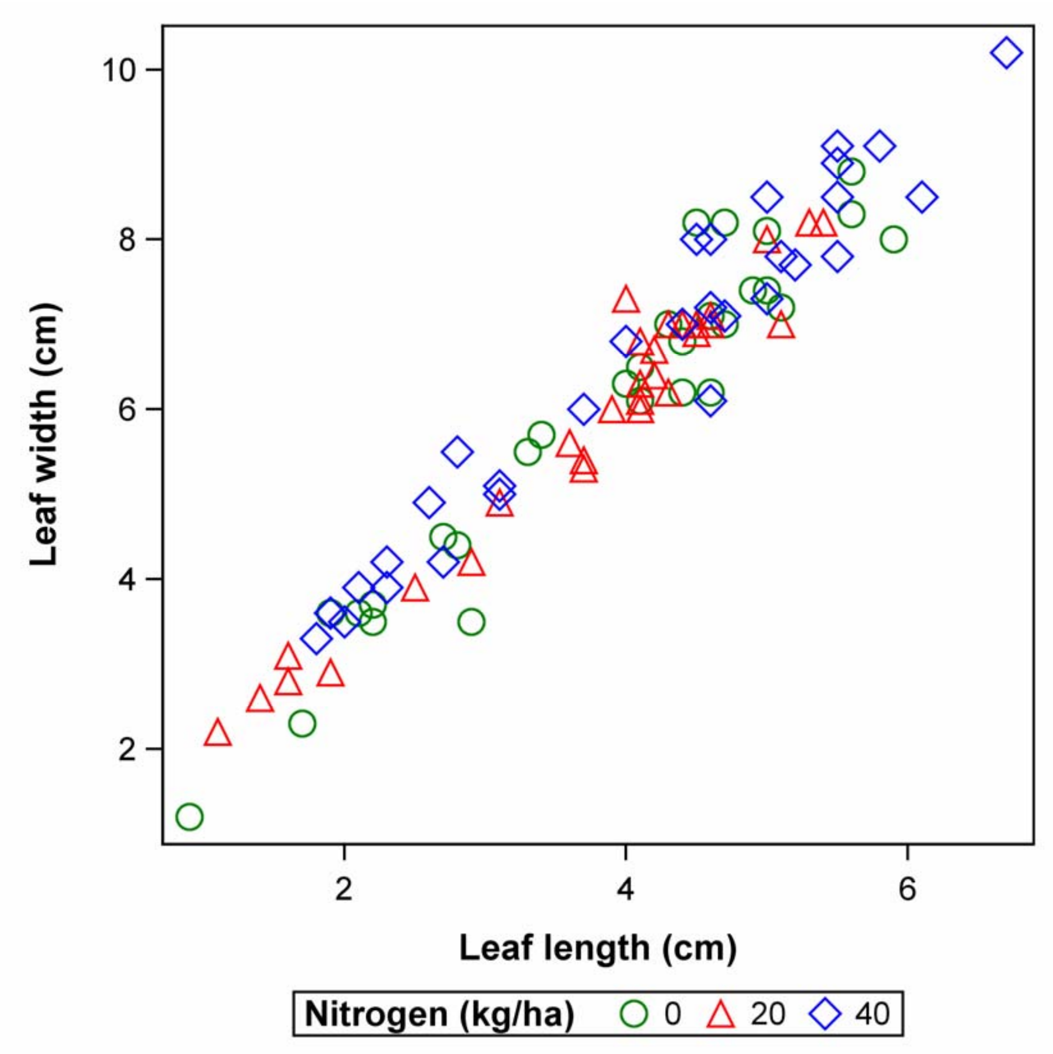
- Chia (*Salvia hispanica* L.) - "superfood"
- Leaf width and leaf length $\Rightarrow$ prediction of LAI

7. Example 4

Design

- Randomized complete block design
- Three fertilizer treatments (0, 20, 40 kg N ha^{-1})
- 10 leaves sampled per plot (sub-samples, pseudo-replicates)

7. Example 4



$r = 0.9697$

7. Example 4

$$y_{ijk} = \mu + b_j + \tau_i + e_{ij} + f_{ijk}$$

where

y_{ijk} = length/width of *k*-th leaf of *i*-th treatment in *j*-th block

b_j = effect of *j*-th block

τ_i = effect of *i*-th treatment

e_{ij} = error of *ij*-th plot

f_{ijk} = error of *ijk*-th measured value y_{ijk}

7. Example 4

Table 7: Correlation estimates of chia data.
Trait 1 = leaf length (cm), trait 2 = leaf width (cm).

Parameter	Estimate	s.e.
$\rho_{b(1,2)}$	-	-
$\rho_{\tau(1,2)}$	1.0000	-
$\rho_{e(1,2)}$	1.0000	-
$\rho_{f(1,2)}$	0.9736	0.005685
$\rho_{y(1,2)}$	0.9686	

r = 0.9697

8. Summary

- Simple correlation r okay for descriptive purposes
- But is superficial and ignores structure of data & experimental design

  $\Rightarrow$ Effecte for blocks and treatments
- Significance test invalid

Solution:

- Inspection of scatter plots should always be the first step
- Account for effects of design (blocks, plots etc.) and treatments
- Bivariate mixed models allow estimating correlation of all effects

  $\Rightarrow$ fit same effects as in univariate analysis

9. Literature

Mack, L., Capezzone, F., Munz, S., Piepho, H.P., Claupein, W., Phillips, T., Graeff-Hönninger, S. (2017): Non-destructive leaf area estimation for chia (*Salvia hispanica* L.). *Agronomy Journal* **109**, 1960-1969.

Piepho, H.P. (2018). Estimating and testing trait correlations in designed experiments using random effects. *Journal of Agricultural Science* **156**, 59-70.

Thanks!