

STAT 516 Lec 11

Analysis of covariance (ANCOVA)

Karl Gregory

2025-04-15

VO₂ max data from Kuehl (2000)

Change in VO₂ max of 12 males (after-minus-before) randomly assigned to two exercise programs (running, step aerobics). Ages recorded.

<i>Group</i>	<i>Age</i>	<i>Change</i>	<i>Group</i>	<i>Age</i>	<i>Change</i>
Aerobic	31	17.05	Running	23	− 0.87
	23	4.96		22	− 10.74
	27	10.40		22	− 3.27
	28	11.05		25	− 1.97
	22	0.26		27	7.50
	24	2.51		20	− 7.25
Mean	25.83	7.71		23.17	− 2.77
Std. Err.	1.40	2.55		1.01	2.54

Source: D. Allen, Exercise Physiology, University of Arizona.

Which exercise program lead to a greater average change in VO₂ max?

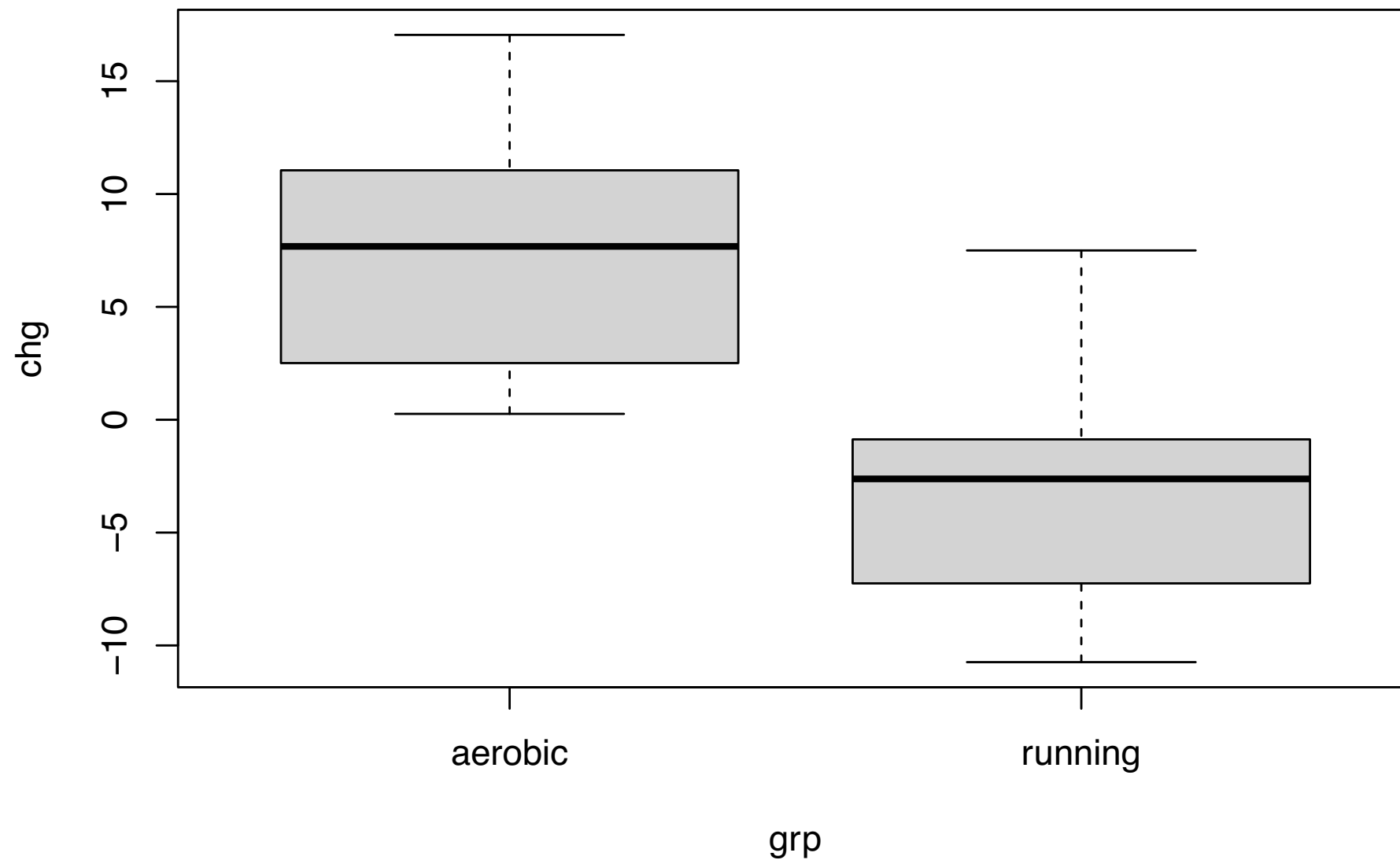
What role does age play?

```
vo2max <- data.frame(chg = c(17.05,4.96,10.40,11.05,0.26,2.51,
                             -0.87,-10.74,-3.27,-1.97,7.50,-7.25),
                    grp = as.factor(c(rep("aerobic",6),rep("running",6))),
                    age = c(31,23,27,28,22,24,23,22,22,25,27,20))
```

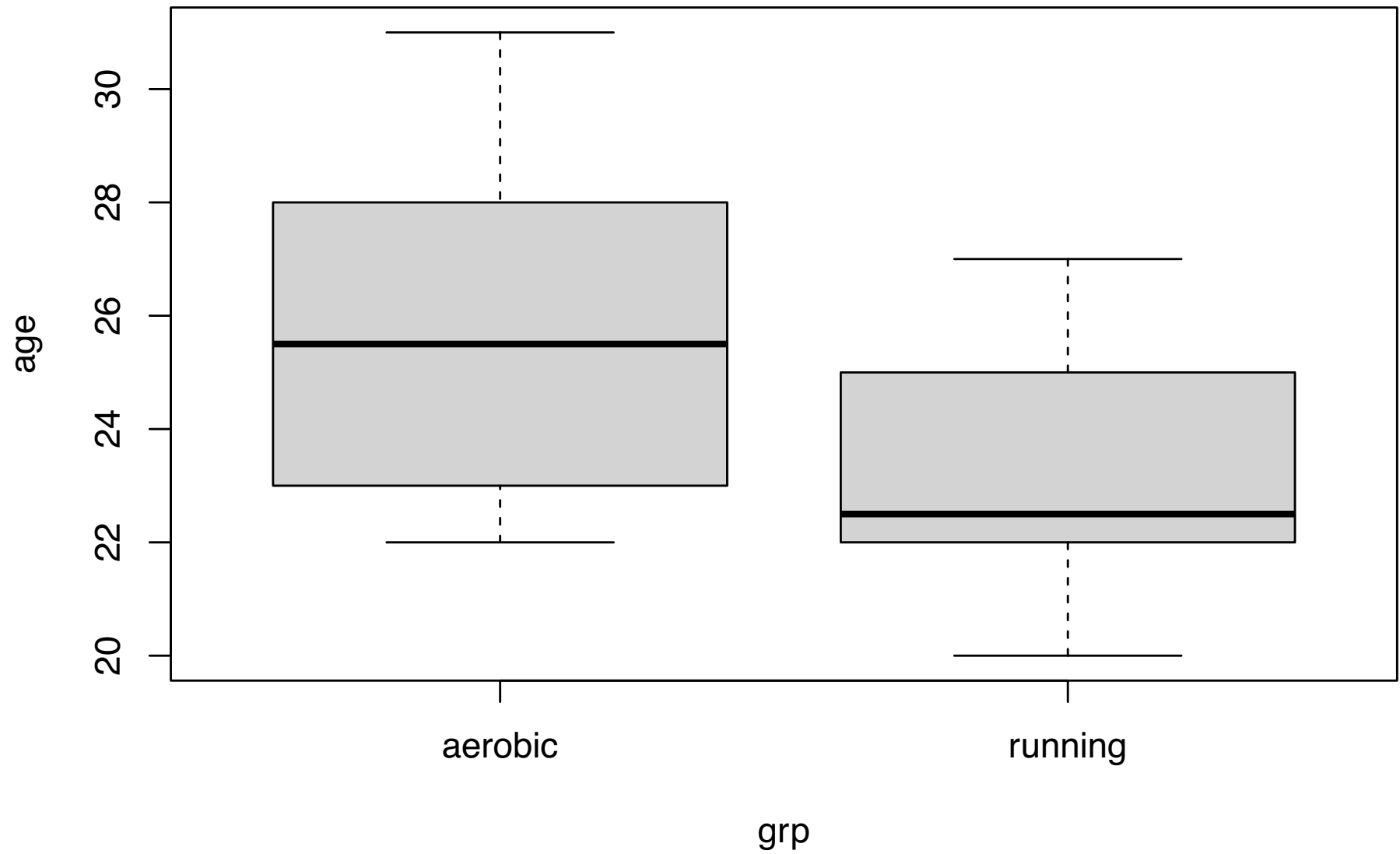
```
vo2max
```

	chg	grp	age
1	17.05	aerobic	31
2	4.96	aerobic	23
3	10.40	aerobic	27
4	11.05	aerobic	28
5	0.26	aerobic	22
6	2.51	aerobic	24
7	-0.87	running	23
8	-10.74	running	22
9	-3.27	running	22
10	-1.97	running	25
11	7.50	running	27
12	-7.25	running	20

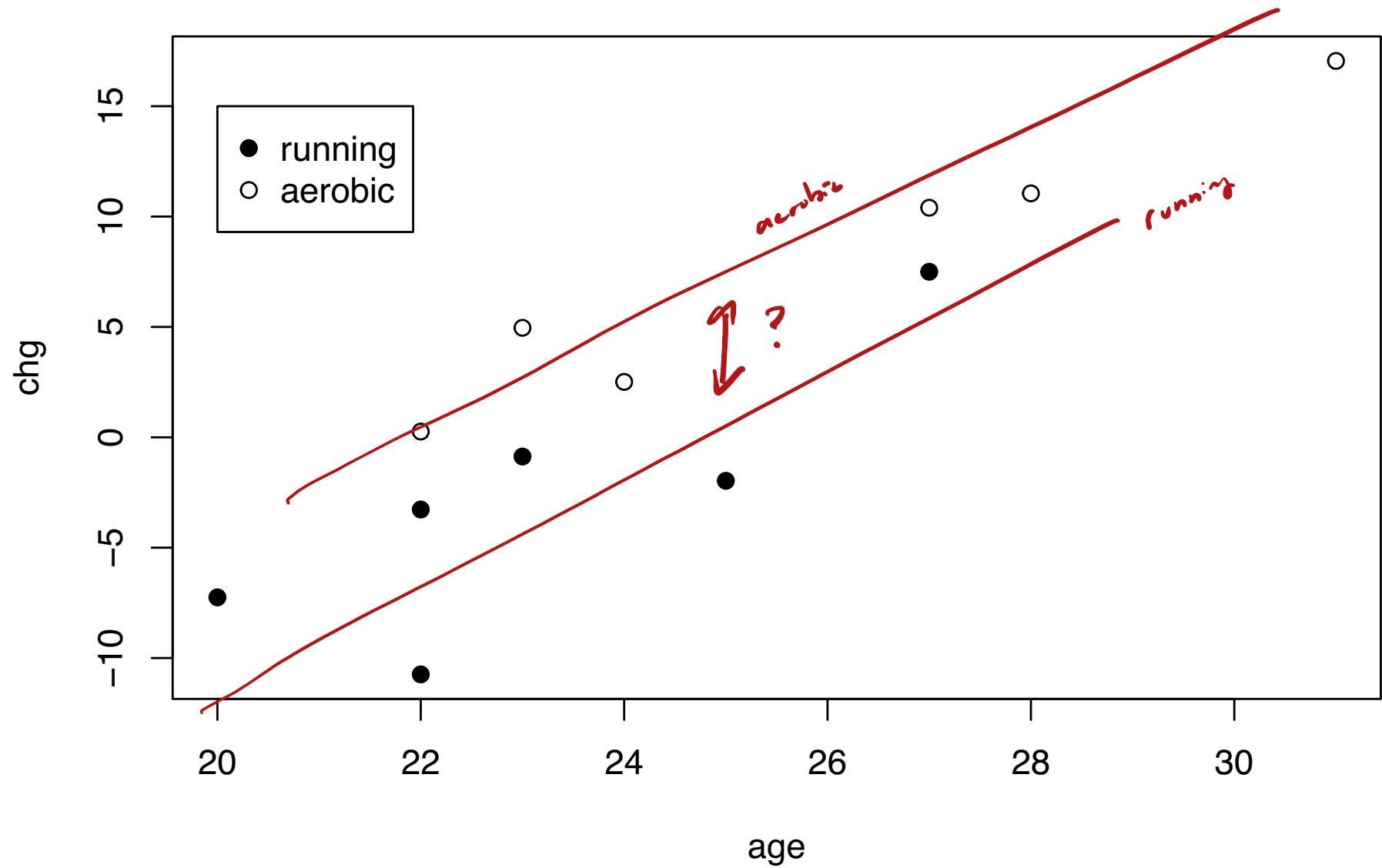
```
boxplot(chg ~ grp, data = vo2max)
```



```
boxplot(age ~ grp, data = vo2max)
```



```
plot(chg ~ age, pch = ifelse(grp == "running",19,1), data = vo2max)
legend(x = 20,y = 15,legend = c("running","aerobic"), pch = c(19,1))
```



Analysis of covariance

subjects of different ages.

age



- ▶ Useful when EUs are not homogeneous.
- ▶ Measurement capturing EU inhomogeneity is called a covariate.
- ▶ Can isolate treatment effects even when EUs differ across treatment groups.

Single-slope analysis of covariance (ANCOVA) model

Assume

for $i = 1, \dots, a$, $j = 1, \dots, n_i$, where

$$Y_{ij} = \mu + \tau_i + \beta x_{ij} + \varepsilon_{ij}$$

a different groups (running vs aerobic, $a=2$)
"slope" for covariate effect
Effect of level i of the factor

the subject/EU

baseline mean

Value of the covariate on EU j in group i .
(age)

noise

- ▶ Y_{ij} is the response of EU j in treatment group i .
- ▶ μ is a baseline or overall mean.
- ▶ the τ_i are treatment effects.
- ▶ the x_{ij} are covariate values measured on the EUs.
- ▶ β is a slope coefficient expressing the effect of the covariate.
- ▶ the ε_{ij} are independent $\text{Normal}(0, \sigma_\varepsilon^2)$ error terms.
- ▶ $n_1 + \dots + n_a = N$. Unbalancedness not an issue.

Set $\mu_i = \mu + \tau_i$ for $i = 1, \dots, a$.

Goals in analysis of covariance

1. Estimate the parameters $\mu, \tau_1, \dots, \tau_a$, and β . lm() function
2. Visualize the data. ← scatter plot
3. Fit full and reduced models and collect error sums of squares.
4. Test whether the covariate has any effect. $H_0: \beta = 0$ vs $H_1: \beta \neq 0$.
5. Test for a treatment effect.
6. Adjust treatment group means for the inhomogeneity of the EUs.
7. Compare the adjusted group means.

Parameter constraints

- ▶ There are a treatment groups and one covariate slope.
- ▶ The model has $a + 1$ parameters, which is one too-many.
- ▶ R will set $\tau_1 = 0$ so that it can estimate all the parameters.

VO₂ max data (cont)

```
lm_out <- lm(chg ~ grp + age, data = vo2max)
summary(lm_out)
```

$$Y_{ij} = \mu + \tau_i + \beta X_{ij} + \epsilon_{ij}$$

$$[R \text{ sets } \tau_i = 0]$$

Call:

```
lm(formula = chg ~ grp + age, data = vo2max)
```

Residuals:

Min	1Q	Median	3Q	Max
-5.7731	-0.9902	0.1395	1.8254	3.0374

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-41.0139	7.7144	-5.317	0.000483 ***
grp _{running}	-5.4426	1.7965	-3.030	0.014255 *
age	1.8859	0.2953	6.386	0.000127 ***

$$\hat{\mu} = -41.0139$$

$$\hat{\tau}_{\text{running}} = -5.4426$$

$$\hat{\beta} = 1.8859$$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.797 on 9 degrees of freedom

Multiple R-squared: 0.902, Adjusted R-squared: 0.8802

F-statistic: 41.42 on 2 and 9 DF, p-value: 2.887e-05

For "aerobic group", the estimated model is
$$-41.0139 + 0 + 1.8859 \times \text{age}$$

for "running group",

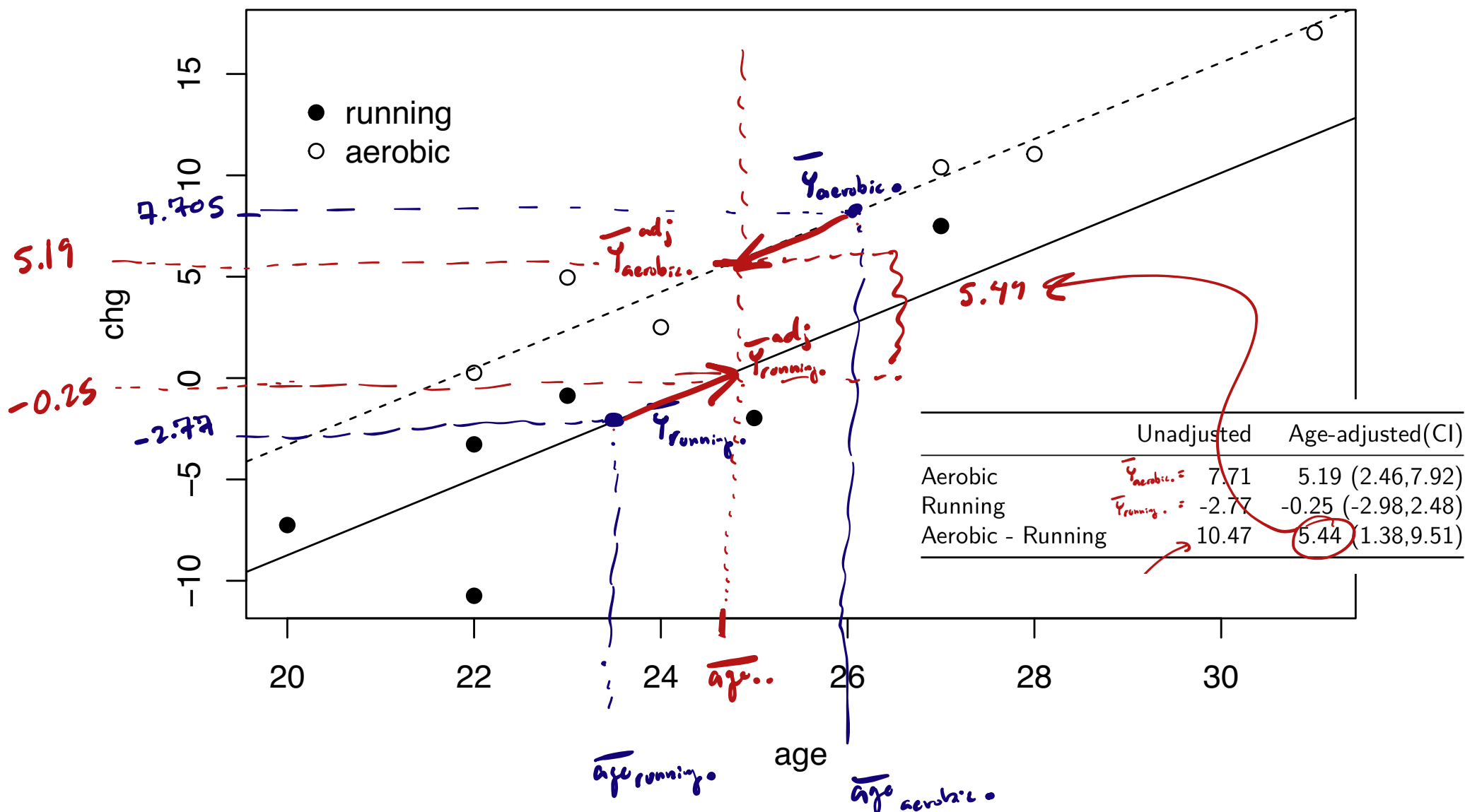
$$-41.0139 + (-5.4426) + 1.8859 \times \text{age}$$

intercept

```
parms <- coef(lm_out)
parms
```

(Intercept)	grprunning	age
-41.013882	-5.442621	1.885892

```
plot(chg ~ age, pch = ifelse(grp == "running",19,1), data = vo2max)
abline(parms[1] + parms[2],parms[3])
abline(parms[1],parms[3],lty = 2)
legend(x = 20,y = 15,legend = c("running","aerobic"), pch = c(19,1), bty = "n")
```



Full and reduced models in ANCOVA

We construct test statistics by fitting full and reduced models:

- ▶ Full model: $Y_{ij} = \mu + \tau_i + \beta x_{ij} + \varepsilon_{ij}$ ← *put everything in.*
- ▶ Reduced model (no covariate effect): $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
- ▶ Reduced model (no treatment effect): $Y_{ij} = \mu + \beta x_{ij} + \varepsilon_{ij}$

Full- and reduced-model error sums of squares

The sum of the squared residuals for each model.

SS

Formula

$SS_{\text{Error}}(\text{Full})$

$$\sum_{i=1}^a \sum_{j=1}^{n_{ij}} (Y_{ij} - (\hat{\mu} + \hat{\tau}_i + \hat{\beta}x_{ij}))^2$$

$SS_{\text{Error}}(\text{No covariate})$

$$\sum_{i=1}^a \sum_{j=1}^{n_{ij}} (Y_{ij} - (\hat{\mu} + \hat{\tau}_i))^2$$

$SS_{\text{Error}}(\text{No treatment})$

$$\sum_{i=1}^a \sum_{j=1}^{n_{ij}} (Y_{ij} - (\hat{\mu} + \hat{\beta}x_{ij}))^2$$

Now define these differences in error sums of squares:

▶ $SS_{\text{Cov}} = SS_{\text{Error}}(\text{No covariate}) - SS_{\text{Error}}(\text{Full})$

▶ $SS_{\text{Trt}} = SS_{\text{Error}}(\text{No treatment}) - SS_{\text{Error}}(\text{Full})$

Covariate sum of squares - Measures importance of covariate
Measures importance of the treatment

ANCOVA table

Analysis of covariance

Source	Df	SS	MS	F value
Covariate	1	SS_{Cov}	MS_{Cov}	$F_{\text{Cov}} = MS_{\text{Cov}} / MS_{\text{Error}}$
Treatment	$a - 1$	SS_{Trt}	MS_{Trt}	$F_{\text{Trt}} = MS_{\text{Trt}} / MS_{\text{Error}}$
Error	$N - a - 1$	SS_{Error}	MS_{Error}	

1. Reject $H_0: \beta = 0$ at α if $F_{\text{Cov}} > F_{1, N-a-1, \alpha}$. *Does covariate have any effect?*
2. Reject $H_0: \mu_1 = \dots = \mu_a$ at α if $F_{\text{Trt}} > F_{a-1, N-a-1, \alpha}$.

$$\mu_i = \mu + \alpha_i$$

All the "intercepts" are the same.

VO₂ max data (cont)

!! Using `anova()` on the `lm()` output gives the wrong SS (gives sequential).!!

Use `Anova()` from R package `car` on the `lm()` output.

```
library(car) # first time run install.packages("car")
# Use type = "II" or type = "III"
Anova(lm_out, type = "III")
```

Anova Table (Type III tests)

Response: chg

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	221.06	1	28.2653	0.0004832 ***
grp	71.79	1	9.1788	0.0142548
age	318.91	1	40.7759	0.0001274 ***
Residuals	70.39	9		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ignore
p-value for testing $H_0: \mu_{\text{running}} = \mu_{\text{aerobics}}$
p-value for testing $H_0: \beta = 0$

```

lm_nocov <- lm(chg ~ grp, data = vo2max) # reduced model with no covariate
lm_notrt <- lm(chg ~ age, data = vo2max) # reduced model with no treatment
lm_full <- lm(chg ~ age + grp, data = vo2max) # full model

SSE_nocov <- sum(lm_nocov$resid^2)
SSE_notrt <- sum(lm_notrt$resid^2)
SSE_full <- sum(lm_full$resid^2)

SSCov <- SSE_nocov - SSE_full
SSTrt <- SSE_notrt - SSE_full
SSE <- SSE_full

a <- 2
N <- nrow(vo2max)

MSCov <- SSCov / 1
MSTrt <- SSTrt / (a-1)
MSE <- SSE / (N - a - 1)

FCov <- MSCov / MSE
FTrt <- MSTrt / MSE

pCov <- 1 - pf(FCov, 1, N - a - 1)
pTrt <- 1 - pf(FTrt, a-1, N - a - 1)

```

Treatment means in terms of parameter estimates

$$\bar{Y}_{aerobic.} = \hat{\mu} + 0 + \hat{\beta}$$

$$\bar{Y}_{running.} = \hat{\mu} + \hat{\tau}_{running} + \hat{\beta}$$

$\bar{x}_{aerobic.}$

$\bar{x}_{running.}$

These means are different.

replace them with the overall average age, $\bar{x}_{..}$

- We can write the treatment group means as

$$\bar{Y}_{i.} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}\bar{x}_{i.} \quad \text{for } i = 1, \dots, a,$$

- We also define the covariate-adjusted means

$$\bar{Y}_{i.}^{\text{adj}} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}\bar{x}_{..} \quad \text{for } i = 1, \dots, a,$$

- We can equivalently write covariate-adjusted means as

$$\bar{Y}_{i.}^{\text{adj}} = \bar{Y}_{i.} - \hat{\beta}(\bar{x}_{i.} - \bar{x}_{..}) \quad \text{for } i = 1, \dots, a.$$

Variances of covariate-adjusted means and their differences

Contrast

Variance

$$\bar{Y}_{i.}^{\text{adj}} \quad \sigma^2 \left[\frac{1}{n_i} + \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{E_{xx}} \right]$$

$$\bar{Y}_{i.}^{\text{adj}} - \bar{Y}_{i'.}^{\text{adj}} \quad \sigma^2 \left[\frac{1}{n_i} + \frac{1}{n_{i'}} + \frac{(\bar{x}_{i.} - \bar{x}_{i'.})^2}{E_{xx}} \right]$$

where $E_{xx} = \sum_{i=1}^a \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2$.

Some (unadjusted) CIs in the ANCOVA model

Define the true or population-level covariate-adjusted means as

$$\mu_i^{\text{adj}} = \mu + \tau_i + \beta \bar{x}_{..} \quad \text{for } i = 1, \dots, a.$$

Target	$(1 - \alpha)100\%$ confidence interval
μ_i^{adj}	$\bar{Y}_{i.}^{\text{adj}} \pm t_{N-a-1, \alpha/2} \sqrt{\text{MS}_{\text{Error}} \sqrt{\frac{1}{n_i} + \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{E_{xx}}}}$
$\mu_i^{\text{adj}} - \mu_{i'}^{\text{adj}}$	$\bar{Y}_{i.}^{\text{adj}} - \bar{Y}_{i'.}^{\text{adj}} \pm t_{N-a-1, \alpha/2} \sqrt{\text{MS}_{\text{Error}} \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}} + \frac{(\bar{x}_{i.} - \bar{x}_{i'.})^2}{E_{xx}}}}$

VO₂ max data (cont)

```
y1. <- mean(vo2max$chg[vo2max$grp == "aerobic"])
y2. <- mean(vo2max$chg[vo2max$grp == "running"])

x1. <- mean(vo2max$age[vo2max$grp == "aerobic"])
x2. <- mean(vo2max$age[vo2max$grp == "running"])
x.. <- mean(vo2max$age)

bhat <- coef(lm_out)[3]

y1.adj <- y1. - bhat * (x1. - x..)
y2.adj <- y2. - bhat * (x2. - x..)
```

```

n1 <- 6
n2 <- 6
alpha <- 0.05

tval <- qt(1-alpha/2,N-a-1)
age1 <- vo2max$age[vo2max$grp == "aerobic"]
age2 <- vo2max$age[vo2max$grp == "running"]
Ex1 <- sum((age1 - mean(age1))^2)
Ex2 <- sum((age2 - mean(age2))^2)
Exx <- Ex1 + Ex2

se1 <- sqrt(MSE) * sqrt(1/n1 + (x1. - x..)^2 / Exx)
lo1 <- y1.adj - tval * se1
up1 <- y1.adj + tval * se1

se2 <- sqrt(MSE) * sqrt(1/n2 + (x2. - x..)^2 / Exx)
lo2 <- y2.adj - tval * se2
up2 <- y2.adj + tval * se2

se12 <- sqrt(MSE) * sqrt(1/n1 + 1/n2 + (x1. - x2.)^2 / Exx)
lo12 <- y1.adj - y2.adj - tval * se12
up12 <- y1.adj - y2.adj + tval * se12

```

Table 5: Mean change in VO_2 max in aerobic and running groups

	Unadjusted	Age-adjusted(CI)
Aerobic	$\bar{Y}_{\text{aerobic}} = 7.71$	5.19 (2.46,7.92)
Running	$\bar{Y}_{\text{running}} = -2.77$	-0.25 (-2.98,2.48)
Aerobic - Running	10.47	5.44 (1.38,9.51)

$$\bar{Y}_{\text{aerobic}} - \bar{Y}_{\text{running}}$$

Soybean data from Dr. Longnecker's notes

Soybean plants assigned to three greenhouse conditions: Supplemental lighting (SL), partial shading (PS), and control (C). The response was seed yield. The pre-treatment height of each plant was also recorded.

Yield	Height	TRT	Yield	Height	TRT	Yield	Height	TRT	Yield	Height	TRT	Yield	Height	TRT
12.2	45	C	12.4	52	C	11.9	42	C	11.3	35	C	11.8	40	C
12.1	48	C	13.1	60	C	12.7	61	C	12.4	50	C	11.4	33	C
12.3	48	C	12.2	51	C	12.6	56	C	13.2	65	C	12.3	51	C
16.6	63	SL	15.8	50	SL	16.5	63	SL	15.0	33	SL	15.4	38	SL
15.6	45	SL	15.8	50	SL	15.8	48	SL	16.0	50	SL	15.8	49	SL
15.0	35	SL	16.2	50	SL	16.7	62	SL	15.8	49	SL	15.9	52	SL
9.5	52	PS	9.5	54	PS	9.6	58	PS	8.8	45	PS	9.5	57	PS
9.8	62	PS	9.1	52	PS	10.3	67	PS	9.5	55	PS	8.5	40	PS
8.6	41	PS	10.4	67	PS	9.4	55	PS	10.2	66	PS	9.3	56	PS

Do the greenhouse conditions effect the seed yield?

What is the role of plant height (proxy for plant vigor)?

```

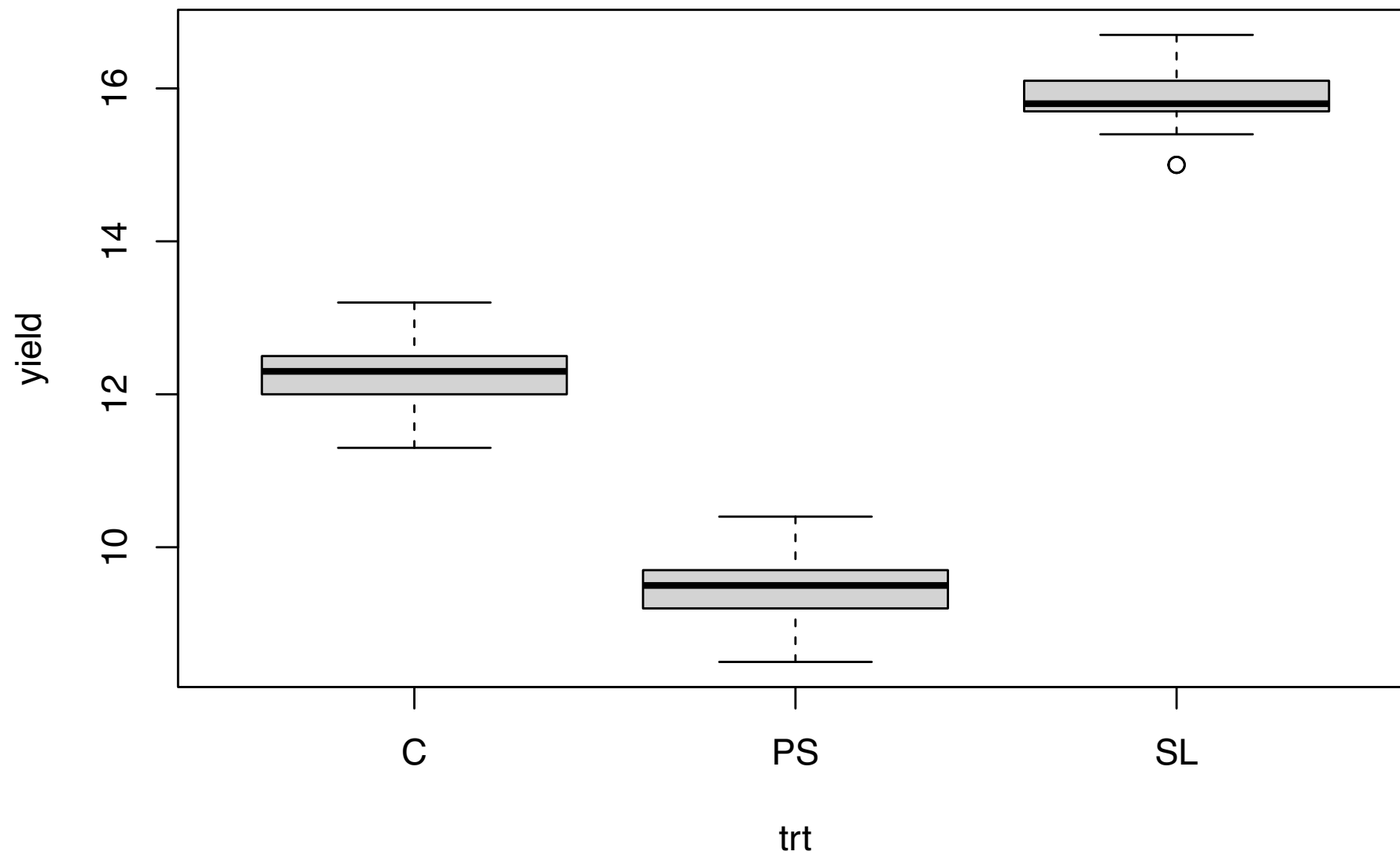
soybean <- data.frame(yield = c(12.2,12.1,12.3,16.6,15.6,15.0,9.5,9.8,8.6,
                                12.4,13.1,12.3,15.8,15.8,16.2,9.5,9.1,10.4,
                                11.9,12.7,12.6,16.5,15.8,16.7,9.6,10.3,9.4,
                                11.3,12.4,13.2,15.0,16.0,15.8,8.8,9.5,10.2,
                                11.8,11.4,12.3,15.4,15.8,15.9,9.5,8.5,9.3),
                      trt = as.factor(rep(c(rep("C",3),rep("SL",3),rep("PS",3)),5)),
                      height = c(45,48,48,63,45,35,52,62,41,
                                 52,60,51,50,50,50,54,52,67,
                                 42,61,56,63,48,62,58,67,55,
                                 35,50,65,33,50,49,45,55,66,
                                 40,33,51,38,49,52,57,40,56))

head(soybean,n = 12)

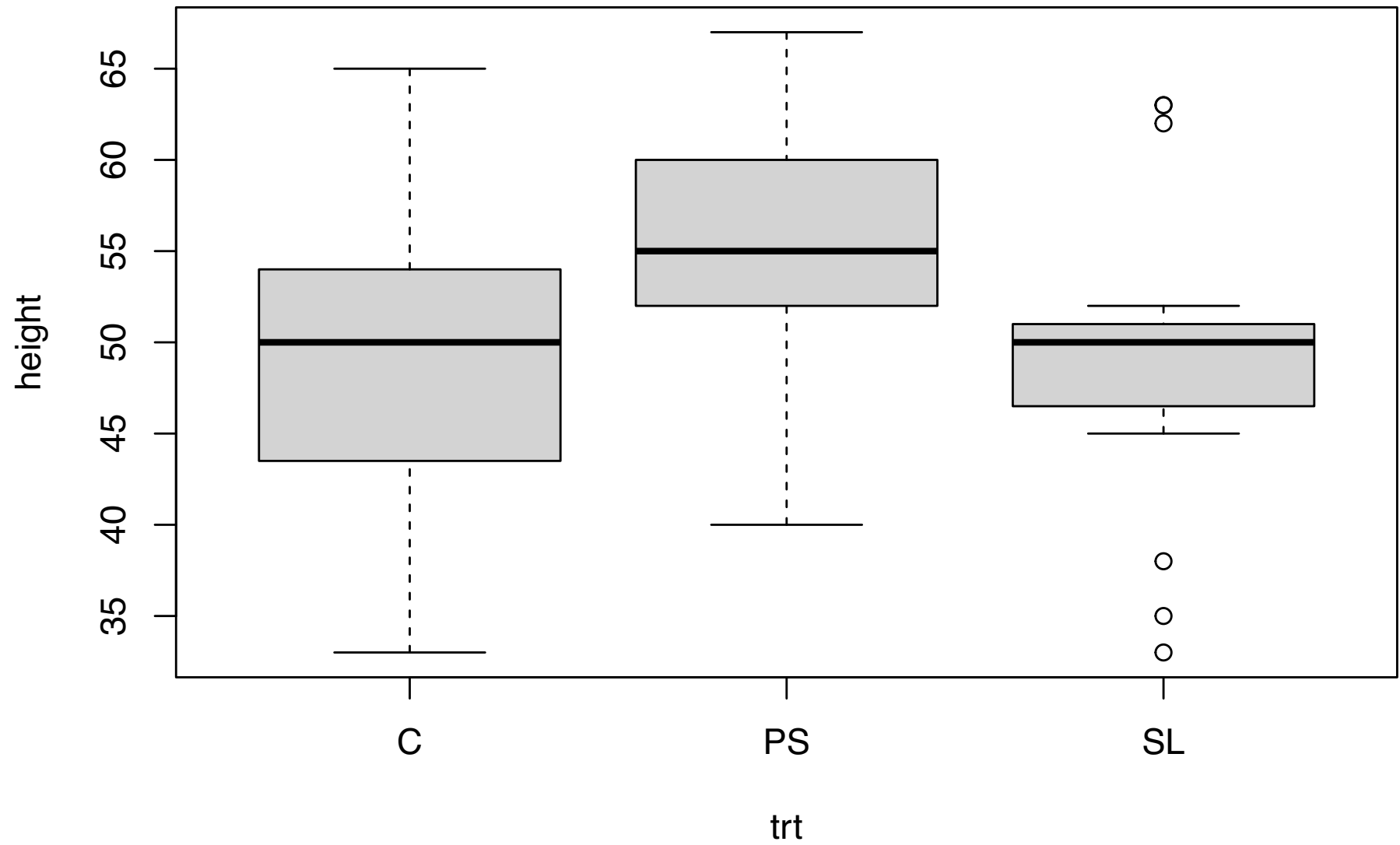
```

	yield	trt	height
1	12.2	C	45
2	12.1	C	48
3	12.3	C	48
4	16.6	SL	63
5	15.6	SL	45
6	15.0	SL	35
7	9.5	PS	52
8	9.8	PS	62
9	8.6	PS	41
10	12.4	C	52
11	13.1	C	60
12	12.3	C	51

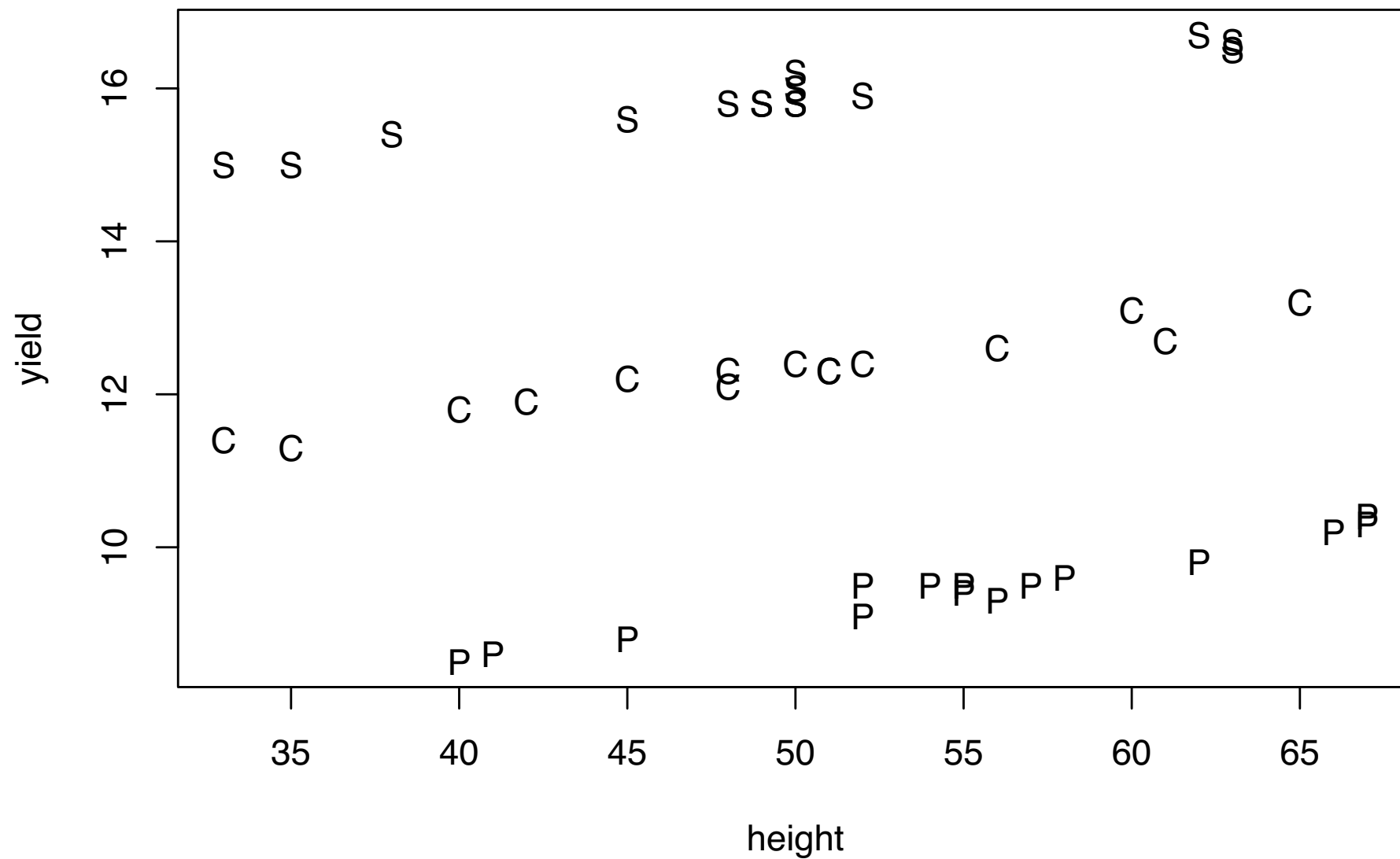
```
boxplot(yield~trt, data = soybean)
```



```
boxplot(height~trt, data = soybean)
```



```
plot(yield ~ height, pch = as.character(trt), data = soybean)
```



ANCOVA allowing different slopes in each group

Previously: $Y_{ij} = \mu + \tau_i + \beta x_{ij} + \varepsilon_{ij}$

Assume

$$Y_{ij} = \mu + \tau_i + (\beta + (\tau\beta)_i) x_{ij} + \varepsilon_{ij}$$

can change the slope for each group.

for $i = 1, \dots, a$, $j = 1, \dots, n_i$, where

- ▶ Y_{ij} is the response of EU j in treatment group i .
- ▶ μ is a baseline or overall mean.
- ▶ the τ_i are treatment effects.
- ▶ the x_{ij} are covariate values measured on the EUs.
- ▶ β is a slope coefficient expressing the effect of the covariate.
- ▶ the $(\tau\beta)_i$ allow interaction between the treatment and the covariate.
- ▶ the ε_{ij} are independent $\text{Normal}(0, \sigma_\varepsilon^2)$ error terms.
- ▶ $n_1 + \dots + n_a = N$. Unbalancedness not an issue.

Set $\mu_i = \mu + \tau_i$ and $\beta_i = \beta + (\tau\beta)_i$ for $i = 1, \dots, a$.

Parameter constraints in multiple slopes model

- ▶ There are a treatment groups and a covariate slopes.
- ▶ The model has $2(a + 1)$ parameters, which is two too-many.
- ▶ R will set $\tau_1 = 0$ and $(\tau\beta)_1 = 0$ to make all parameters estimable.

```
lm_out <- lm(yield ~ trt + height + trt:height, data = soybean)
summary(lm_out)
```

Call:

```
lm(formula = yield ~ trt + height + trt:height, data = soybean)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.234733	-0.088745	-0.003954	0.057644	0.293320

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.500573	0.181934	52.220	< 2e-16 ***
trtPS	-3.639588	0.285527	-12.747	1.74e-15 ***
trtSL	3.713050	0.258575	14.360	< 2e-16 ***
height	0.056298	0.003644	15.451	< 2e-16 ***
trtPS:height	0.009102	0.005372	1.694	0.0982 .
trtSL:height	-0.002437	0.005179	-0.470	0.6407

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1256 on 39 degrees of freedom

Multiple R-squared: 0.9981, Adjusted R-squared: 0.9978

F-statistic: 4054 on 5 and 39 DF, p-value: < 2.2e-16

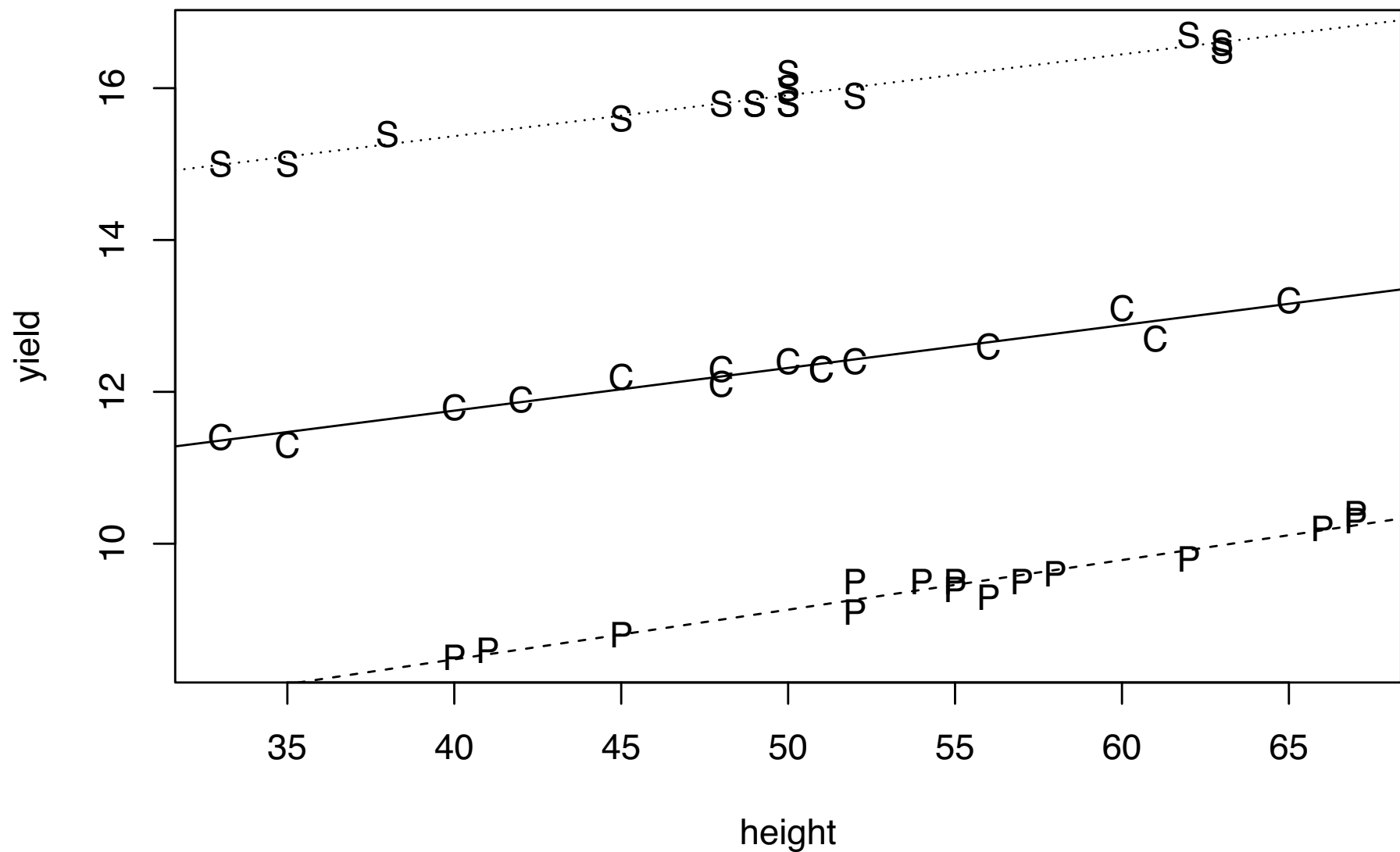
```
parms <- coef(lm_out)
parms
```

(Intercept)	trtPS	trtSL	height	trtPS:height	trtSL:height
9.500572519	-3.639588070	3.713050304	0.056297710	0.009101605	-0.002436573

```

plot(yield ~ height, pch = as.character(trt), data = soybean)
abline(parms[1],parms[4],lty = 1)
abline(parms[1] + parms[2],parms[4] + parms[5],lty = 2)
abline(parms[1] + parms[3],parms[4] + parms[6],lty = 3)

```



An F-test for equal slopes

Define the error sums of squares:

SS	Formula
$SS_{\text{Error}}(\text{Full})$	$\sum_{i=1}^a \sum_{j=1}^{n_{ij}} (Y_{ij} - (\hat{\mu} + \hat{\tau}_i + (\hat{\beta} + (\hat{\tau}\hat{\beta})_i)x_{ij}))^2$
$SS_{\text{Error}}(\text{Equal slopes})$	$\sum_{i=1}^a \sum_{j=1}^{n_{ij}} (Y_{ij} - (\hat{\mu} + \hat{\tau}_i + \hat{\beta}x_{ij}))^2$

allows different slopes

forces equal slopes.

Now set

$$F_{T \times C} = \frac{[SS_{\text{Error}}(\text{Equal slopes}) - SS_{\text{Error}}(\text{Full})]/(a - 1)}{SS_{\text{Error}}(\text{Full})/(N - 2a)}$$

Reject H_0 : *Equal slopes* at α if $F_{T \times C} > F_{a-1, N-2a, \alpha}$.

Soybean data (cont)

```
lm_eqslp <- lm(yield ~ trt + height, data = soybean) # equal slopes model
lm_full <- lm(yield ~ trt + height + trt:height, data = soybean) # full model

SSE_eqslp <- sum(lm_eqslp$resid^2)
SSE_full <- sum(lm_full$resid^2)

a <- 3
N <- nrow(soybean)

FTrtCov <- ((SSE_eqslp - SSE_full) / (a-1)) / (SSE_full / (N - 2*a))
pTrtCov <- 1 - pf(FTrtCov, a-1, N - 2*a)
```

Same as interaction p value from `anova()` on the `lm()` output.

```
anova(lm_out)
```

Analysis of Variance Table

Response: yield

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
trt	2	308.124	154.067	9770.7982	< 2e-16	***
height	1	11.389	11.389	722.2854	< 2e-16	***
trt:height	2	0.079	0.039	2.4938	0.09568	.
Residuals	39	0.615	0.016			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

p-value for testing
 H_0 : equal slopes.

When slopes are unequal

If we reject H_0 : *Equal slopes*, then

- ▶ We can compute covariate-adjusted means with the different slopes.
- ▶ We can compare the covariate adjusted means—but the CI formulas are different from the ones above (details omitted).

References

Kuehl, R. O. 2000. *Design of Experiments: Statistical Principles of Research Design and Analysis*. Duxbury/Thomson Learning.