

## ANCOVA

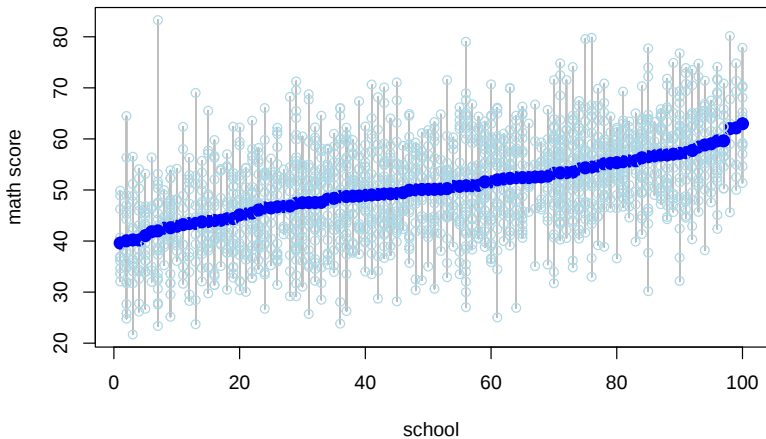
Peter Hoff  
Duke STA 610

## Motivating example

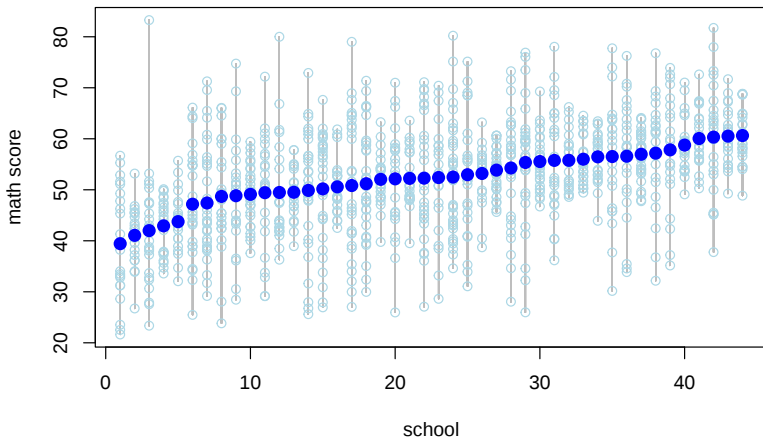
## ANCOVA

## NELS analysis

## NELS data



## Heteroscedasticity



## Heteroscedasticity

Levene's test: If  $\sigma_i^2$  is large, then  $|y_{i,j} - \bar{y}_j| = |\hat{\epsilon}_{i,j}|$  should be large.

- Let  $z_{i,j} = |\hat{\epsilon}_{i,j}|$
- Use the ANOVA  $F$ -test for across-group differences *in the  $z_{i,j}$ 's*

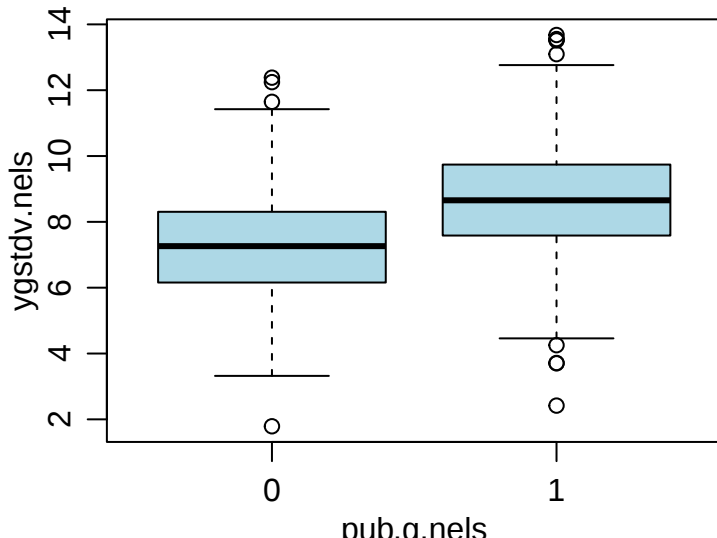
```
fit.nels<-lm(y.nels~as.factor(g.nels))
z.nels<-abs( fit.nels$res )
anova(lm(z.nels~as.factor(g.nels)) )

## Analysis of Variance Table
##
## Response: z.nels
##
```

|                   | Df    | Sum Sq | Mean Sq | F value | Pr(>F)        |
|-------------------|-------|--------|---------|---------|---------------|
| as.factor(g.nels) | 683   | 27078  | 39.645  | 1.6092  | < 2.2e-16 *** |
| Residuals         | 12290 | 302776 | 24.636  |         |               |

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

What kind of schools might have higher variation?



## Within-group variance models

**Homoscedastic model:**  $y_{i,j} \sim N(\theta_j, \sigma^2)$ .

- Simple to implement;
- The estimate of  $\sigma^2$  will be precise if assumption is correct;
- The assumption could be wrong!

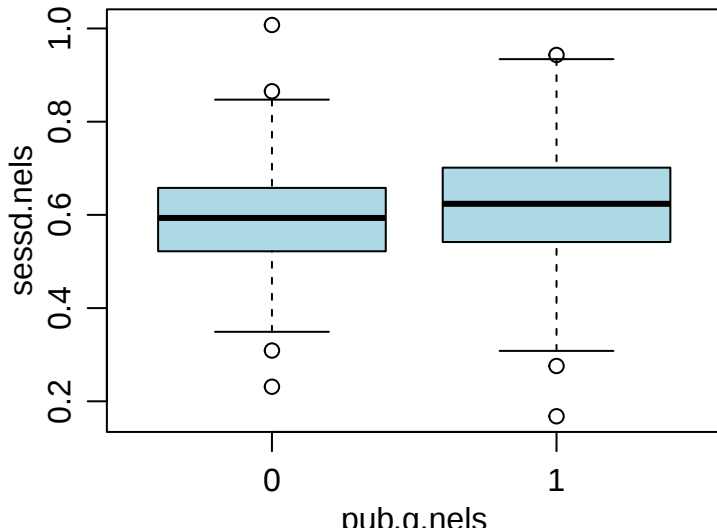
**Heteroscedastic hierarchical normal model:**

- Use  $\hat{\sigma}_j^2 = \sum (y_{i,j} - \bar{y}_j)^2 / (n_j - 1)$  if  $n_j$ 's are large.
- Alternatively, use a hierarchical model for the variances.
- More appropriate inferences if variances are truly different.

But this doesn't explain *why* variances are different.

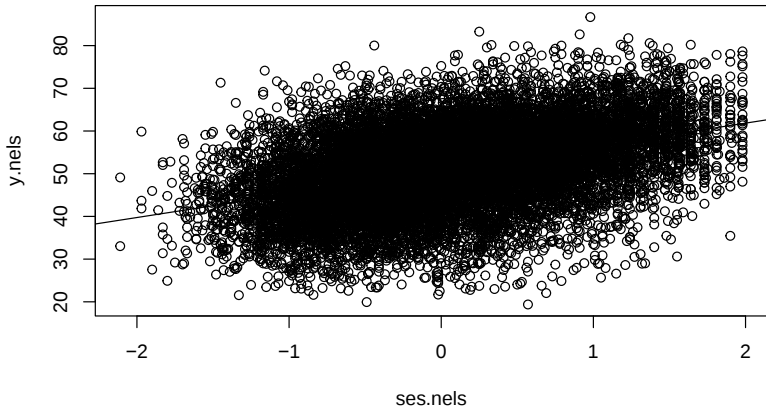
**Variance due to observable factors:**

- Outcome could be related to unit-level characteristics  $x_{i,j}$ ;
- Within-group variance can be partitioned:
  - variance explainable by observable unit-level characteristics;
  - unexplained variation.

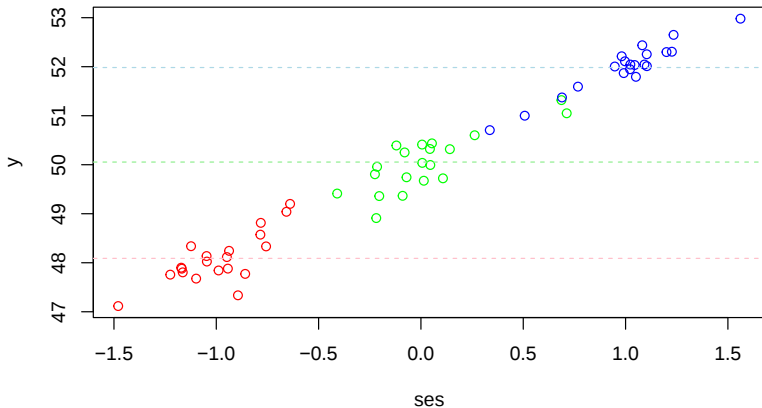


## Marginal relationship

```
plot(y.nels~ses.nels)
abline(lm(y.nels~ses.nels))
```

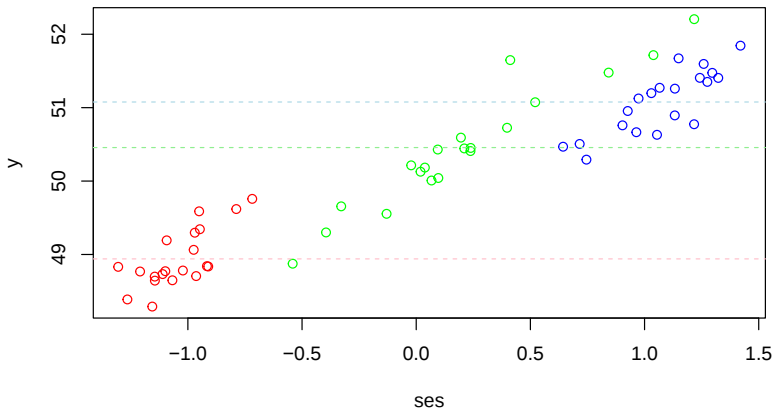


## Possible explanations



### Variation across schools attributable to student-level variation in SES

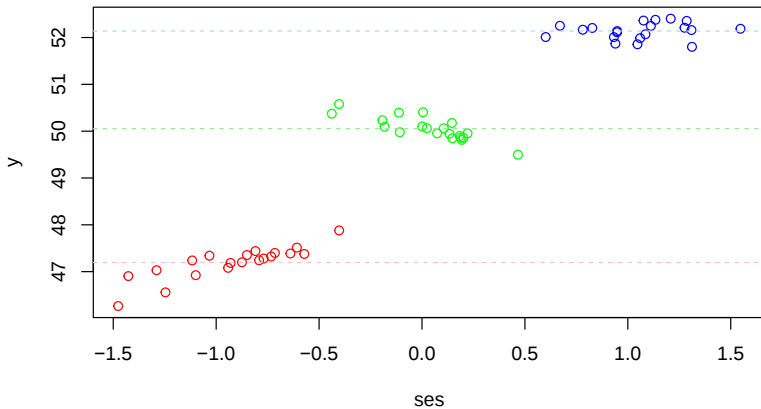
## Possible explanations



## Variance across schools partially attributable to student-level variation in SES

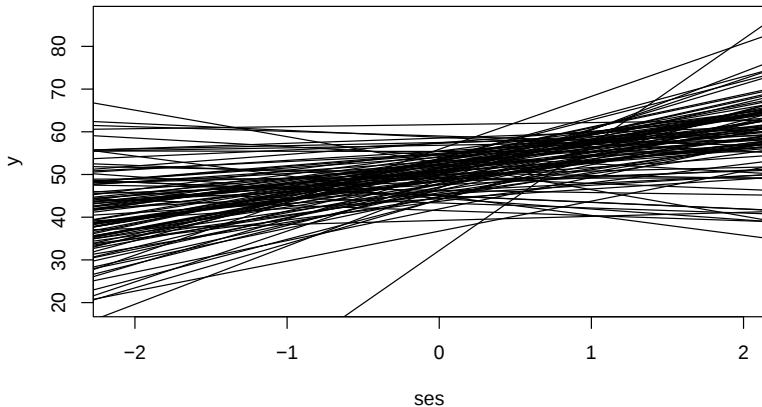
Variance across schools not attributable to student-level variation in SES

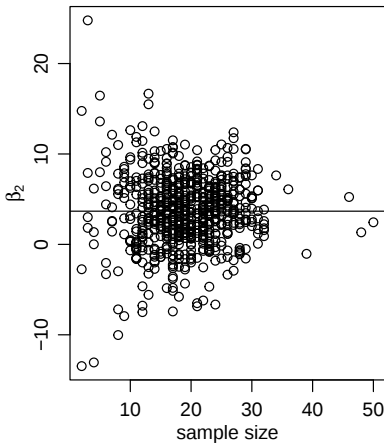
## Possible explanations



## School specific OLS estimates

$$y_{i,j} = \hat{\beta}_{1,j} + \hat{\beta}_{2,j}x_{i,j} + \epsilon_{i,j}$$





## Estimation and testing

### Hierarchical approach:

$$\begin{aligned} y_{i,j} &= \beta_{1,j} + \beta_{2,j}x_{i,j} + \epsilon_{i,j} \\ &= (\beta_1 + a_{1,j}) + (\beta_2 + a_{2,j})x_{i,j} + \epsilon_{i,j}, \end{aligned}$$

### Testing:

- Do the  $a_{1,j}$ 's vary across groups?  $H_0 : a_{1,j} = 0$  for all  $j$ .
- Do the  $a_{2,j}$ 's vary across groups?  $H_0 : a_{2,j} = 0$  for all  $j$ .

Note if  $a_{1,j} = a_{1,j'} = 0$  for all  $j$ , then

- There still may be real heterogeneity in *mean* test scores, but
- all heterogeneity is attributable to heterogeneity in  $x_{i,j}$ .

**Estimation:** If  $H_0$  is rejected, how do we estimate  $\beta_{1,j}, \beta_{2,j}$ ?

- Unbiased OLS estimates?
- Biased shrinkage estimates?

**Task:** Estimate  $p(y|\mathbf{x})$  from the data.

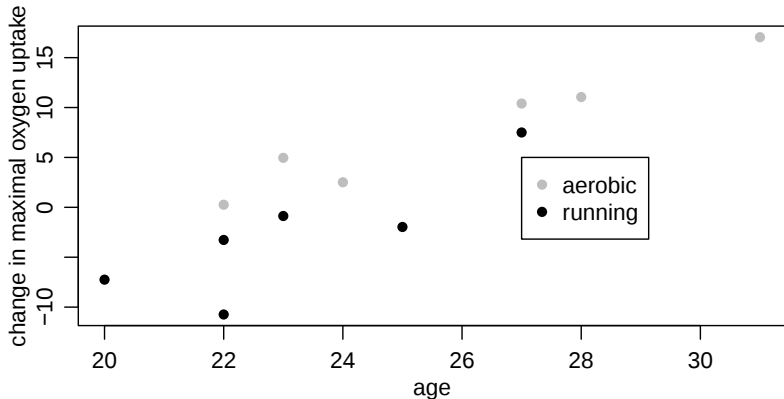
### Example: O<sub>2</sub> uptake

**Study design:** 12 men randomly assigned to one of two regimens:

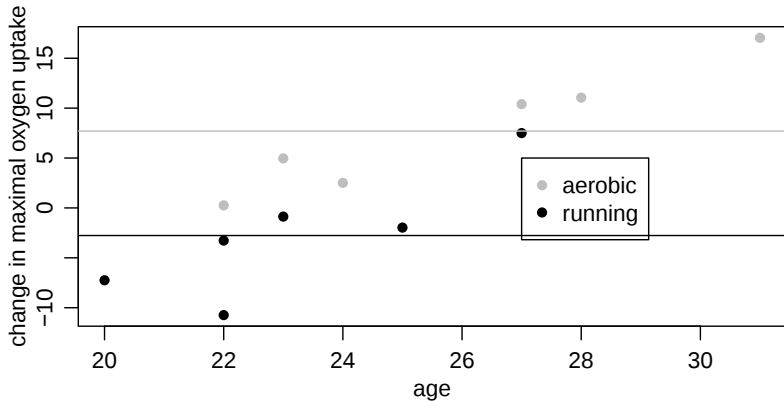
- flat terrain running;
- step aerobics.

The maximal O<sub>2</sub> uptake of each participant was measured after 3 months.

Age data is also available.



### Example: O<sub>2</sub> uptake



```
mean(y[aerobic==1])
## [1] 7.705
mean(y[aerobic==0])
## [1] -2.766667
```



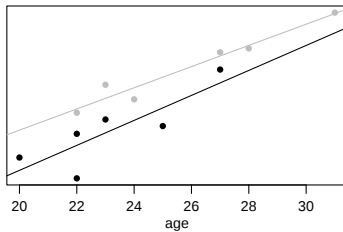
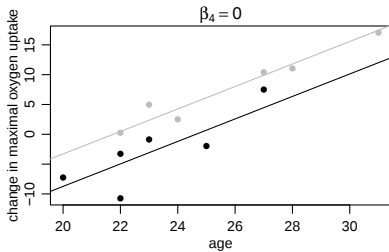
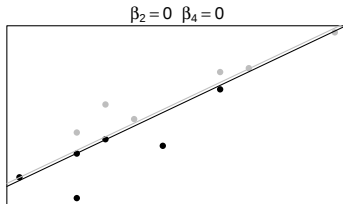
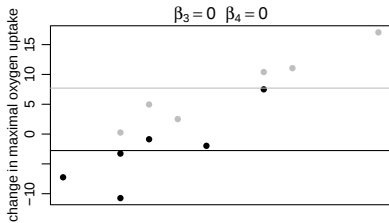
## Linear regression for $O_2$ uptake

$$\begin{aligned} y_i &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i, \text{ where} \\ x_{i,1} &= 1 \text{ for each subject } i \\ x_{i,2} &= 0 \text{ if subject } i \text{ is on the running program, 1 if on aerobic} \\ x_{i,3} &= \text{age of subject } i \\ x_{i,4} &= x_{i,2} \times x_{i,3} \end{aligned}$$

The conditional expectations of  $y$  for the two levels of  $x_{j,2}$  are models as

$$\begin{aligned} E[y|x] &= \beta_1 + \beta_3 \times \text{age} && \text{if on running program} \\ E[y|x] &= (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \text{age} && \text{if on aerobic program} \end{aligned}$$

## Submodels



## Normal linear regression

A full statistical model requires

- A specification of  $E[y|\mathbf{x}]$  (the “mean model”)
- A specification of the distribution of  $y$  around  $E[y|\mathbf{x}]$

**Normal linear regression:**

$$y_i = \boldsymbol{\beta}^T \mathbf{x}_i + \epsilon_i$$

$$\epsilon_1, \dots, \epsilon_n \sim \text{i.i.d. normal}(0, \sigma^2)$$

**Vector-matrix form:** Let  $\mathbf{y}$  be the  $n$ -dimensional column vector  $(y_1, \dots, y_n)^T$ , and  $\mathbf{X}$  be the  $n \times p$  matrix with  $i$ th row  $\mathbf{x}_i$ . The normal regression model is

$$\{\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2\} \sim \text{multivariate normal } (\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}),$$

where  $\mathbf{I}$  is the  $p \times p$  identity matrix and

$$\mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \mathbf{x}_1 \rightarrow \\ \mathbf{x}_2 \rightarrow \\ \vdots \\ \mathbf{x}_n \rightarrow \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \dots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \dots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} E[y_1|\boldsymbol{\beta}, \mathbf{x}_1] \\ \vdots \\ E[y_n|\boldsymbol{\beta}, \mathbf{x}_n] \end{pmatrix}.$$

## OLS estimation

For any given value of  $\beta$ ,

- the fitted value for observation  $i$  is  $\beta^T \mathbf{x}_i$ ;
- the error or residual for  $i$  is  $(y_i - \beta^T \mathbf{x}_i)$  ;
- the SSE for  $\beta$  is

$$\begin{aligned} \text{SSE}(\beta) &= \sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2 \\ &= \|\mathbf{y} - \mathbf{X}\beta\|^2. \end{aligned}$$

The *ordinary least-squares* (OLS) estimate  $\hat{\beta}$  of  $\beta$  minimizes SSE.

## OLS regression

To find the minimizing value of  $\beta$ , rewrite  $SSE(\beta)$  in matrix notation:

$$\begin{aligned} SSE(\beta) &= \sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta \end{aligned}$$

Recall from calculus that

1. a minimum of a function  $g(z)$  occurs at a value  $z$  such that  $\frac{d}{dz}g(z) = 0$ ;
2. the derivative of  $g(z) = az$  is  $a$  and the derivative of  $g(z) = bz^2$  is  $2bz$ .

## OLS estimation

$$\begin{aligned}
 \frac{d}{d\beta} \text{SSE}(\beta) &= \frac{d}{d\beta} \left( \mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta \right) \\
 &= -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \beta, \text{ therefore} \\
 \frac{d}{d\beta} \text{SSE}(\beta) = 0 &\Leftrightarrow -2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \beta = 0 \\
 &\Leftrightarrow \mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y} \\
 &\Leftrightarrow \beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}
 \end{aligned}$$

$\hat{\beta}_{ols} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$  is OLS estimate of  $\beta$ .

## OLS estimation for the O<sub>2</sub> uptake data

```
X
##      int trt age trt.age
## [1,]   1   0  23       0
## [2,]   1   0  22       0
## [3,]   1   0  22       0
## [4,]   1   0  25       0
## [5,]   1   0  27       0
## [6,]   1   0  20       0
## [7,]   1   1  31      31
## [8,]   1   1  23      23
## [9,]   1   1  27      27
## [10,]  1   1  28      28
## [11,]  1   1  22      22
## [12,]  1   1  24      24

y
## [1] -0.87 -10.74 -3.27 -1.97  7.50 -7.25 17.05  4.96 10.40 11.05
## [11]  0.26  2.51
```

## OLS estimation for the O<sub>2</sub> uptake data

```
XtX<-t(X)%*%X
```

```
XtX
```

```
##          int trt  age trt.age
## int      12   6  294    155
## trt       6   6  155    155
## age      294 155 7314   4063
## trt.age  155 155 4063   4063
```

```
Xty<-t(X)%*%y
```

```
Xty
```

```
##          [,1]
## int      29.63
## trt      46.23
## age     978.81
## trt.age 1298.79
```

```
solve(XtX) %*% Xty
```

```
##          [,1]
## int     -51.2939459
## trt      13.1070904
## age       2.0947027
## trt.age  -0.3182438
```

## OLS estimation for the O<sub>2</sub> uptake data

```

solve(XtX) %*% Xty

##               [,1]
## int      -51.2939459
## trt       13.1070904
## age        2.0947027
## trt.age   -0.3182438

# with indicators
aerobic

## [1] 0 0 0 0 0 0 1 1 1 1 1 1

lm(y~aerobic+age+aerobic*age)

##
## Call:
## lm(formula = y ~ aerobic + age + aerobic * age)
##
## Coefficients:
## (Intercept)      aerobic          age  aerobic:age
##      -51.2939      13.1071       2.0947       -0.3182

```

## OLS estimation for the O<sub>2</sub> uptake data

```
# with factors
trt

## [1] "running" "running" "running" "running" "running" "running" "running" "aerobic"
## [8] "aerobic" "aerobic" "aerobic" "aerobic" "aerobic"

fit<-lm(y~trt+age+trt*age)

# aerobic is baseline
fit

##
## Call:
## lm(formula = y ~ trt + age + trt * age)
##
## Coefficients:
##      (Intercept)      trtrunning          age  trtrunning:age
##      -38.1869      -13.1071          1.7765          0.3182

fit$coef[1]+fit$coef[2]

## (Intercept)
##   -51.29395

fit$coef[3]+fit$coef[4]

##      age
## 2.094703
```

## Properties of OLS estimates

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

**Unbiasedness:** Treating  $\mathbf{X}$  as fixed for the moment,

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}] \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E[\mathbf{y}] \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

**Variance:** Conditional on  $\mathbf{X}$ ,

$$\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

## Optimality of OLS

**UMVUE:** If  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$   $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$  then

$$\text{Var}[\hat{\boldsymbol{\beta}}] < \text{Var}[\tilde{\boldsymbol{\beta}}]$$

for any other *unbiased* estimator  $\tilde{\boldsymbol{\beta}}$ .

**BLUE:** If  $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$ ,  $\text{Var}[\mathbf{y}|\mathbf{X}] = \sigma^2 \mathbf{I}$  then

$$\text{Var}[\hat{\boldsymbol{\beta}}] < \text{Var}[\tilde{\boldsymbol{\beta}}]$$

for any other *linear unbiased* estimator  $\tilde{\boldsymbol{\beta}}$ , that is

- $\tilde{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y}$  for some  $\mathbf{A} \in \mathbb{R}^{p \times n}$ ;
- $E[\tilde{\boldsymbol{\beta}}|\mathbf{X}, \boldsymbol{\beta}] = \boldsymbol{\beta}$  for all  $\boldsymbol{\beta}$ ;

## Standard errors and CIs

$$\epsilon_1, \dots, \epsilon_n \sim \text{iid } N(0, \sigma^2)$$

How can we estimate  $\sigma^2$ ?

**Idea:** Since  $\beta \approx \hat{\beta}$ ,

$$\begin{aligned}\epsilon_i &= y_i - \beta^T \mathbf{x}_i \\ &\approx y_i - \hat{\beta}^T \mathbf{x}_i = \hat{\epsilon}_i\end{aligned}$$

$$\text{sample variance}(\epsilon_1, \dots, \epsilon_n) \approx \sigma^2$$

$$\text{sample variance}(\hat{\epsilon}_1, \dots, \hat{\epsilon}_n) \approx \sigma^2$$

**SSE:** Let  $SSE = \sum (y_i - \hat{\beta}^T \mathbf{x}_i)^2 = \sum \hat{\epsilon}_i^2$ .

$$\hat{\sigma}^2 = \frac{SSE}{n - p} \quad (\text{unbiased estimator})$$

$$\hat{\sigma}^2 = \frac{SSE}{n} \quad (\text{maximum likelihood estimator})$$

## Variance-covariance for the O<sub>2</sub> uptake data

```
beta.ols<-solve(XtX) %*% Xty
res<- y-XtX%*beta.ols

SSE<-sum(res^2)

s2.hat<-SSE/( length(res) - length(beta.ols) )

VB<-s2.hat* solve(XtX)
```

VB

```
##          int          trt          age      trt.age
## int      150.116712 -150.116712 -6.4184014   6.4184014
## trt     -150.116712  248.439893  6.4184014 -10.1693473
## age      -6.418401   6.418401  0.2770533  -0.2770533
## trt.age   6.418401 -10.169347 -0.2770533   0.4222512
```

```
sqrt(diag(VB))
```

```
##          int          trt          age      trt.age
## 12.2522126 15.7619762  0.5263585   0.6498086
```

## Variance-covariance for the O<sub>2</sub> uptake data

```
fit<-lm(y~aerobic+age+aerobic*age)
summary(fit)

##
## Call:
## lm(formula = y ~ aerobic + age + aerobic * age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.5295 -0.9610  0.3945  2.1717  2.2883
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -51.2939     12.2522  -4.187  0.00305 **
## aerobic      13.1071     15.7620   0.832  0.42978
## age          2.0947      0.5264   3.980  0.00406 **
## aerobic:age  -0.3182      0.6498  -0.490  0.63746
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.923 on 8 degrees of freedom
## Multiple R-squared:  0.9049, Adjusted R-squared:  0.8692
## F-statistic: 25.36 on 3 and 8 DF,  p-value: 0.0001938

beta.ols/sqrt(diag(VB))

##              [,1]
## int      -4.1865047
## trt       0.8315639
## age       3.9796120
## trt.age  -0.4897500
```

## Evaluating group effects, the ANCOVA view

**ANOVA:** Evaluate heterogeneity across categorical factors with an  $F$ -test.

**ANCOVA:** Evaluate heterogeneity across categorical factors with an  $F$ -test, *after accounting for a (continuous) covariate.*

### Questions answered:

- ANOVA: is there heterogeneity across groups?
- ANCOVA: is there heterogeneity across groups, *beyond that attributable to a covariate* ?

## Standard ANCOVA model

$$y_{i,j} = (\beta_0 + b_{0,j}) + \beta_1 \times x_{i,j} + \epsilon_{i,j}$$

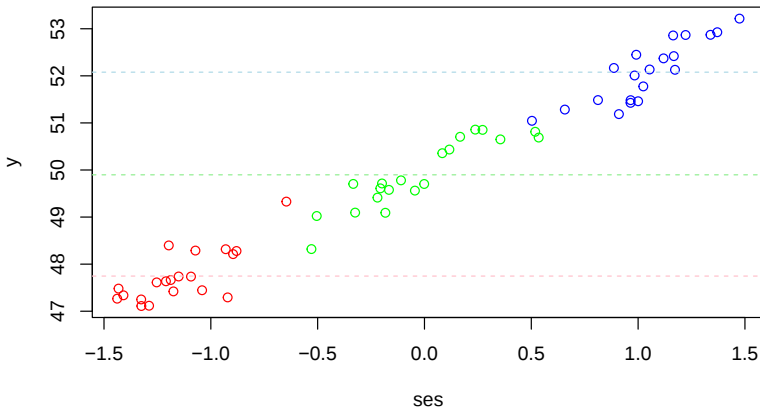
- $y_{i,j}$  refers to the  $i$ th observation in group  $j$ ;
- $b_{0,j}$  refers to the effect of  $j$ th group on the mean;
- $\beta_1$  refers to the slope (assumed identical across groups).

For two-groups the model is the same as the following regression model:

$$y_i = (\beta_0 + b_0 \times \text{aerobic}_i) + \beta_1 \times \text{age} + \epsilon_i$$

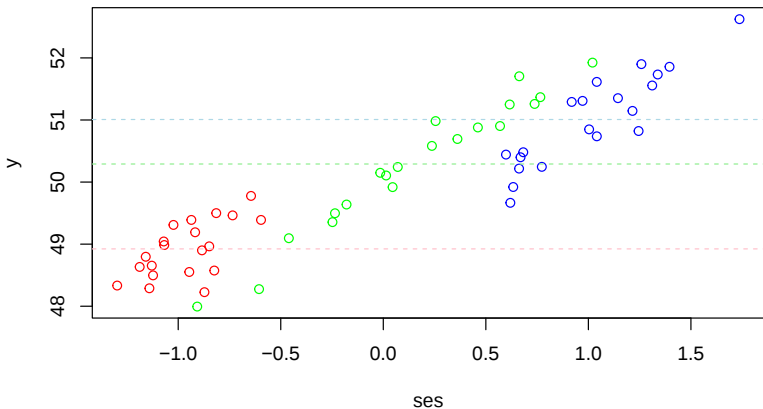
- $y_i$  is the  $i$ th observation overall;
- $\text{aerobic}_i$  is the indicator that person  $i$  is in the aerobics group;

## Possible explanations



$$b_0 = 0.$$

## Possible explanations



$$b_0 \neq 0.$$

## Testing and ANCOVA

$$y_{i,j} = (\beta_0 + b_{0,j}) + \beta x_{i,j} + \epsilon_{i,j}$$

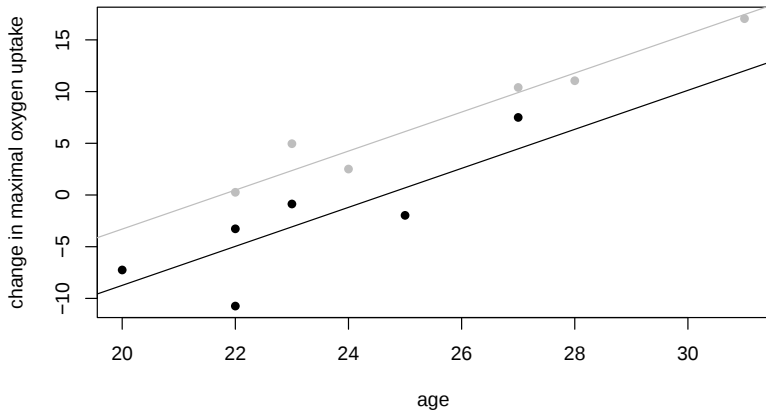
A test of across-group heterogeneity is provided by an  $F$ -test:

```
fit1<-lm( y~ age + as.factor(trt))
anova(fit1)

## Analysis of Variance Table
##
## Response: y
##          Df Sum Sq Mean Sq F value    Pr(>F)
## age       1  576.09   576.09  73.6594 1.257e-05 ***
## as.factor(trt) 1   71.79    71.79   9.1788 0.01425 *
## Residuals    9   70.39     7.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The  $p$ -value indicates evidence of across-group heterogeneity beyond that attributable to age.

## Variable intercept model



## ANCOVA with interactions

$$y_{i,j} = (\beta_0 + b_{0,j}) + (\beta_1 + b_{1,j})x_{i,j} + \epsilon_{i,j}$$

- $b_{1,j}$  is a group specific slope parameter

For two-groups the model is the same as the following regression model:

$$y_i = (\beta_0 + b_0 \times \text{aerobic}_i) + (\beta_1 + b_1 \times \text{aerobic}_i) \times \text{age}_i + \epsilon_i$$

- $\text{aerobic}_i$  is the indicator that person  $i$  is in the aerobics group;
- $b_1$  is the difference in slopes between the two groups.

## ANCOVA with interactions

```
fit2<-lm( y~ age + as.factor(trt) + age*as.factor(trt) )
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: y
##
```

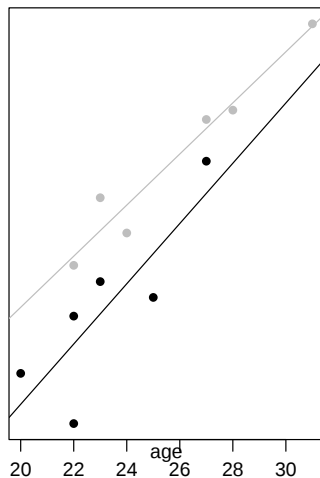
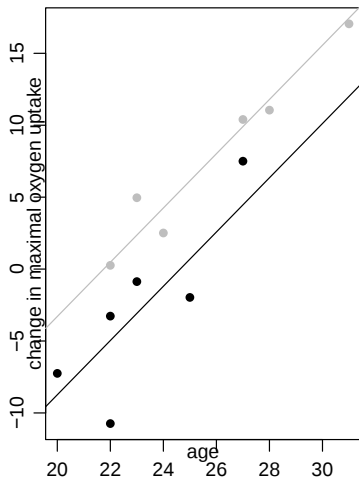
|                       | Df | Sum Sq | Mean Sq | F value | Pr(>F)    |     |
|-----------------------|----|--------|---------|---------|-----------|-----|
| ## age                | 1  | 576.09 | 576.09  | 67.4381 | 3.615e-05 | *** |
| ## as.factor(trt)     | 1  | 71.79  | 71.79   | 8.4035  | 0.01993   | *   |
| ## age:as.factor(trt) | 1  | 2.05   | 2.05    | 0.2399  | 0.63746   |     |
| ## Residuals          | 8  | 68.34  | 8.54    |         |           |     |

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

There is not evidence for heterogeneity beyond what can be attributed to

- age
- a mean difference between groups

## ANCOVA with interactions



## Heterogeneous regressions

It will be convenient to rewrite the model in vector form:

$$y_{i,j} = \beta_j^T \mathbf{x}_{i,j} + \epsilon_{i,j}$$

$$\beta_j = \beta + \mathbf{b}_j$$

- $\beta$  represents the average across-group relationship of  $y$  to  $\mathbf{x}$ .
- $\{\mathbf{b}_1, \dots, \mathbf{b}_m\}$  represent across-group heterogeneity of the relationship.

In the O<sub>2</sub> uptake example,

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \mathbf{b}_j = \begin{pmatrix} b_{0,j} \\ b_{1,j} \end{pmatrix} \quad \mathbf{x}_{i,j} = \begin{pmatrix} 1 \\ \text{age}_{i,j} \end{pmatrix}$$

$$\begin{aligned} E[y_{i,j}] &= \beta_j^T \mathbf{x}_{i,j} = [\beta + \mathbf{b}_j]^T \mathbf{x}_{i,j} \\ &= \beta^T \mathbf{x}_{i,j} + \mathbf{b}_j^T \mathbf{x}_{i,j} \\ &= [\beta_0 + \beta_1 \times \text{age}_{i,j}] + [b_{0,j} + b_{1,j} \times \text{age}_{i,j}] \end{aligned}$$

## Testing for an overall group effect

Sometimes it will be more convenient to test for *any* group effect:

$$y_{i,j} = \boldsymbol{\beta}^T \mathbf{x}_{i,j} + \mathbf{b}_j^T \mathbf{x}_{i,j} + \epsilon_{i,j}$$

$$H_0: \mathbf{b}_1 = \dots = \mathbf{b}_m = \mathbf{0}$$

$$H_1: \mathbf{b}_j \neq \mathbf{0}, \text{ some } j \in \{1, \dots, m\}$$

This can be done via an  $F$ -test as well:

```
fit0<-lm( y~ age )
fit1<-lm( y~ age + as.factor(trt) )
fit2<-lm( y~ age + as.factor(trt) + age*as.factor(trt) )
```

## Testing for an overall group effect

```
anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: y
##              Df Sum Sq Mean Sq F value    Pr(>F)
## age              1  576.09   576.09  67.4381 3.615e-05 ***
## as.factor(trt)    1   71.79    71.79   8.4035  0.01993 *
## age:as.factor(trt) 1    2.05     2.05   0.2399  0.63746
## Residuals        8   68.34     8.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit0,fit2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ age
## Model 2: y ~ age + as.factor(trt) + age * as.factor(trt)
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      10 142.18
## 2       8  68.34  2    73.836 4.3217 0.05338 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Why overall tests?

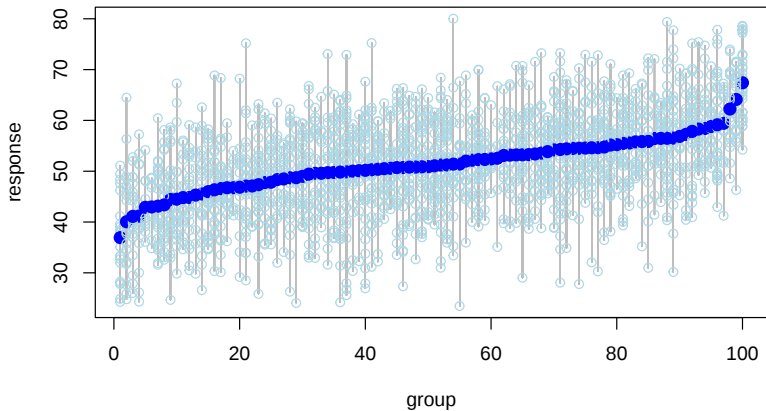
Consider a scenario where we have lots of regressors:

$$\begin{aligned} y_{i,j} &= \beta_j^T \mathbf{x}_{i,j} + \epsilon_{i,j} \\ &= \beta_{1,j} x_{1,i,j} + \cdots + \beta_{p,j} x_{p,i,j} + \epsilon_{i,j} \end{aligned}$$

Compare and contrast the following two procedures:

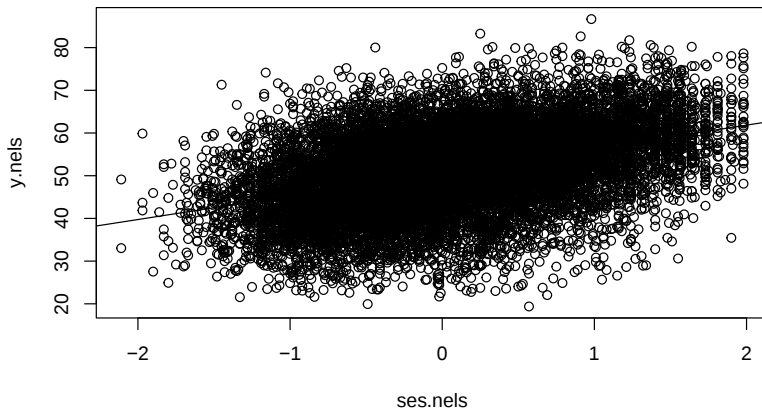
1. Iteratively search for predictors that show across group heterogeneity;
2. Perform an overall test of across-group differences
  - If heterogeneity detected, describe it for each predictor.

## NELS data



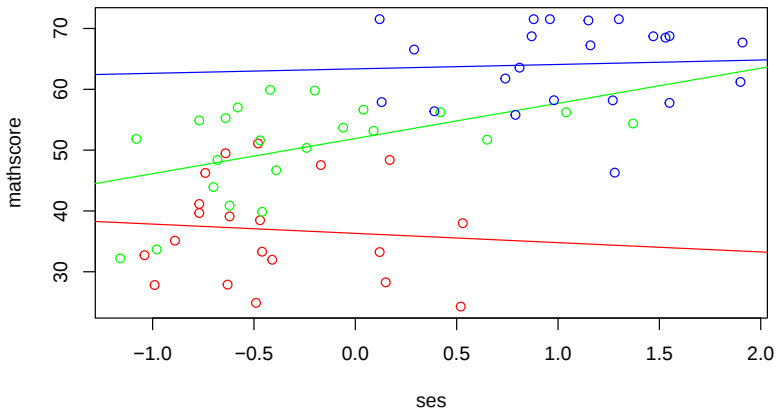
## Marginal relationship

```
plot(y.nels~ses.nels)
abline(lm(y.nels~ses.nels))
```



- Micro effects of SES on mathscore;
- Macro effects of SES on mathscore.

## Some actual data



What explanations do these data support?

## OLS approach

```
BETA<-NULL
for(j in sort(unique(g.nels)))
{
  yj<-y.nels[g.nels==j]
  xj<-ses.nels[g.nels==j]
  fitj<-lm(yj~xj)
  BETA<-rbind(BETA,fitj$coef)
}
```

### some results

```
BETA[1:10,]
```

```
##      (Intercept)      xj
## [1,]    53.02066  5.0815402
## [2,]    49.82444  2.9045055
## [3,]    38.48130  1.1340111
## [4,]    46.38335  2.6715294
## [5,]    46.35686  5.0231028
## [6,]    48.96969  0.9272974
## [7,]    46.26290  6.8041213
## [8,]    53.39039  5.0407659
## [9,]    51.73138  2.5813744
## [10,]   49.84851  4.9972552
```

```
### mean intercept, mean slope
apply(BETA, 2, mean, na.rm=TRUE)

## (Intercept)          xj
## 50.618228      3.672483

### compare to pooled analysis
lm(y.nels~ses.nels)

##
## Call:
## lm(formula = y.nels ~ ses.nels)
##
## Coefficients:
## (Intercept)      ses.nels
##      50.793         5.527
```

What does the discrepancy suggest in terms of macro vs micro effects of SES?



## Testing for heterogeneity

```
### sequential test of effects
anova(fit1)

## Analysis of Variance Table
##
## Response: y.nels
##
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## ses.nels      1 223914   223914  3347.0036 < 2.2e-16 ***
## as.factor(g.nels) 683 190150      278    4.1615 < 2.2e-16 ***
## ses.nels:as.factor(g.nels) 682  56264      82    1.2332 4.865e-05 ***
## Residuals    11607 776507      67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

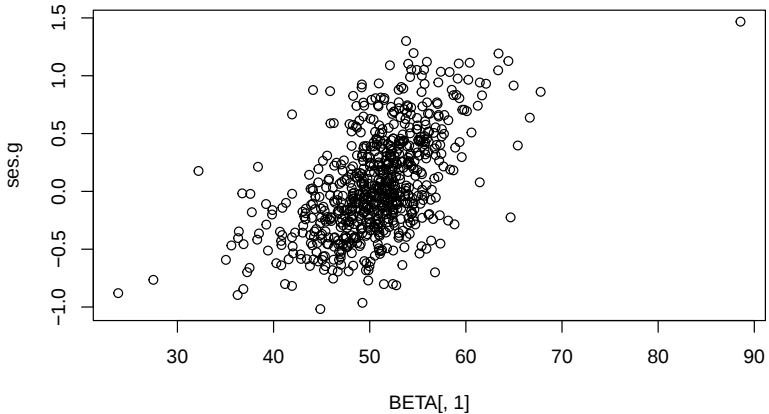
The data provide strong evidence of across-group heterogeneity in mathscore/SES association.

Furthermore, the data suggest both

- micro-level effects of SES (slopes are on average positive)
- macro-level effects of SES (average slope is lower than pooled slope)

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

## Macro-level effects



$$\mathbf{x}_j^T \mathbf{x}_j = \sum_{i=1}^{n_j} \mathbf{x}_{i,j} \mathbf{x}_{i,j}^T \text{ is generally increasing in } n_j$$