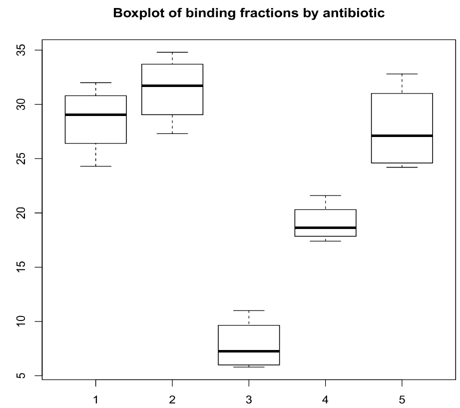
Single factor experiments

Following data come from study investigating binding fraction for several antibiotics using $n = 20$ bovine serum samples:



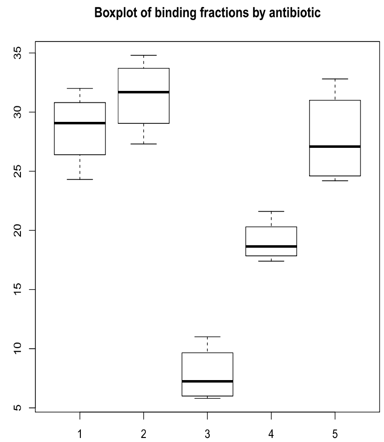
Antibiotic	Binding Percentage				Sample mean	Sample s.d.
Penicillin G	29.6	24.3	28.5	32	28.6	3.2
Tetracyclin	27.3	32.6	30.8	34.8	31.4	3.2
Streptomycin	5.8	6.2	11	8.3	7.8	2.4
Erythromycin	21.6	17.4	18.3	19	19.1	1.8
Chloramphenicol	29.2	32.8	25	24.2	27.8	4.0

A ____________________________ (CRD) was used.

(All assignment of antibiotics to serum samples equally likely.)

Q: Are the population means for these 5 treatments plausibly equal?

Q: Do these (sample) treatment means differ significantly?




```
> bf.dat <- read.table("bindingfractions.txt",header=T)
> with(bf.dat,
+ boxplot(y~drug,main="Boxplot of binding fractions by
+ antibiotic"))
> dev.copy2pdf(file="bindingfractions.pdf")
X11cairo
       2
```

```
> bf.dat$drug <- as.factor(bf.dat$drug)
> bf.out <- lm(bf.dat$y ~ bf.dat$drug)
> anova(bf.out)
Analysis of Variance Table

Response: bf.dat$y
            Df  Sum Sq Mean Sq F value    Pr(>F)
bf.dat$drug  4 1480.82  370.21  40.885 6.74e-08 ***
Residuals   15  135.82    9.05
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
```

An ANOVA table. Note that $df = 4$.

Modelling the binding fraction expt

"Effects" model parameterizes antibiotic effects as differences from mean:

$$Y_{ij} = \mu + \tau_i + E_{ij}$$

for $i = 1, \ldots, 5$ and $j = 1, \ldots, 4$, where E_{ij} are i.i.d. $N(0, \sigma^2)$ errors.

- μ - overall population mean (avg of 5 treatment population means)
- τ_i - difference between (population) mean for treatment i and μ
- σ^2 - (population) variance of bf for a given antibiotic

To test H_0 : ______________________ $= 0$, we just carry out one-way ANOVA:

Source	d.f.	Sum of squares	Mean Square	F	p-value
Treatments	4	1481	370	41	
Error	15	136	9		
Total	19	1617			

Conclusion? (Use $F(0.05, 4, 15) = 3.06$)

Parameter estimates $\hat{\mu}, \hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4, \hat{\tau}_5$? *Six parameters. Not uniquely estimable*

$$\widehat{\mu + \tau_1} = \overline{y}_{1\cdot} = 28.6, \ldots, \widehat{\mu + \tau_5} = \overline{y}_{5\cdot} = 27.8$$

Standard errors of parameter estimates?

$\text{StdErr}(\widehat{\mu + \tau_i}) = \text{StdErr}(\overline{y}_{i\cdot}) =$ ______________

Table for balanced one-way ANOVA

Y_{ij} denotes j^{th} observation receiving level i of treatment factor with t levels, for a total of N observations.

Source	d.f.	Sum of squares	Mean Square	F
Treatments	$t-1$	$SS[T]$	$MS[T] = \frac{SS[T]}{(t-1)}$	$F = \frac{MS[T]}{MS[E]}$
Error	$N-t$	$SS[E]$	$MS[E] = \frac{SS[E]}{(N-t)}$	
Total	$N-1$	$SS[TOT]$		

where

$$\begin{aligned} SS[T] &= \sum\sum(\bar{y}_{i+} - \bar{y}_{..})^2 \\ SS[E] &= \sum\sum(y_{ij} - \bar{y}_{i+})^2 \\ SS[TOT] &= \sum\sum(y_{ij} - \bar{y}_{..})^2 \end{aligned}$$

The linear model $\mu_{ij} = E(Y_{ij}) = \mu + \tau_i$ could be fit using MLR with 5 **indicator variables** $x_1, \ldots, x_5$ for the 5 antibiotics. Let

$$x_{ij} = \begin{cases} 1 & \text{if treatment } j \\ 0 & \text{else} \end{cases}$$

A general linear model

Models which parameterize the effects of classification factors this way are general linear models. One-way ANOVA and linear regression models are general linear models. The linearity pertains to the parameters, not the explanatory variables.

Here, reparameterizing using $5 - 1$ indicator variables leads to a general linear model. Define x_1, x_2, x_3, x_4 as before. Then the MLR model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + E_i \quad i = 1, \ldots, 20$$

where $E_i \overset{iid}{\sim} N(0, \sigma^2)$.
The X matrix looks like

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Remarks:

- $(X'X)^{-1}$ exists
- continuously valued covariates (as opposed to indicators) can be added and it is still a general linear model

For the one-way ANOVA,

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} 27.8 \\ 0.8 \\ 3.6 \\ -20.0 \\ -8.7 \end{pmatrix}$$

Estimates for the five treatment means obtained by substitution of $\hat{\beta}$ into $\mu(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

$$\begin{aligned}
\hat{\mu}(1,0,0,0) &= (1,1,0,0,0)\widehat{\beta} &= \hat{\beta}_0 + \hat{\beta}_1 &= 28.6 \\
\hat{\mu}(0,1,0,0) &= (1,0,1,0,0)\widehat{\beta} &= \hat{\beta}_0 + \hat{\beta}_2 &= 31.4 \\
\hat{\mu}(0,0,1,0) &= (1,0,0,1,0)\widehat{\beta} &= \hat{\beta}_0 + \hat{\beta}_3 &= 7.8 \\
\hat{\mu}(0,0,0,1) &= (1,0,0,0,1)\widehat{\beta} &= \underline{\hspace{6cm}} &= \underline{\hspace{2cm}} \\
\hat{\mu}(0,0,0,0) &= (1,0,0,0,0)\widehat{\beta} &= \underline{\hspace{6cm}} &= 27.8
\end{aligned}$$

(Compare with slide 1.)

For standard errors, use $\hat{\Sigma}$:

$$\hat{\Sigma} = MS[E](X'X)^{-1} = \begin{pmatrix} 2.26 & -2.26 & -2.26 & -2.26 & -2.26 \\ & 4.53 & 2.26 & 2.26 & 2.26 \\ & & 4.53 & 2.26 & 2.26 \\ & & & 4.53 & 2.26 \\ & & & & 4.53 \end{pmatrix}$$

Let a, b, c, d be defined by

$$a' = (1,1,0,0,0), b' = (1,0,1,0,0), c' = (1,0,0,1,0), d' = (1,0,0,0,1).$$

Then

$$\begin{array}{rclcl} \hat{\mu}(1,0,0,0) & = & \hat{\beta}_0 + \hat{\beta}_1 & = & a'\hat{\beta} \\ \hat{\mu}(0,1,0,0) & = & \hat{\beta}_0 + \hat{\beta}_2 & = & b'\hat{\beta} \\ \hat{\mu}(0,0,1,0) & = & \hat{\beta}_0 + \hat{\beta}_3 & = & c'\hat{\beta} \\ \hat{\mu}(0,0,0,1) & = & \hat{\beta}_0 + \hat{\beta}_4 & = & d'\hat{\beta} \\ \hat{\mu}(0,0,0,0) & = & \hat{\beta}_0 & = & \hat{\beta}_0 \end{array}$$

and

$$a'\hat{\Sigma}a = b'\hat{\Sigma}b = c'\hat{\Sigma}c = d'\hat{\Sigma}d = \hat{\Sigma}_{11} = 2.3 = \widehat{\text{Var}}(\hat{\beta}_0) = \widehat{\text{Var}}(\hat{\beta}_0 + \hat{\beta}_j)$$

so the estimated SE for any sample treatment mean is $\sqrt{2.3} = 1.5$.

Checking matrix arithmetic:

$$
\begin{aligned}
(1,1,0,0,0)\widehat{\Sigma}\begin{pmatrix}1\\1\\0\\0\\0\end{pmatrix} &= \\
&= (1,1,0,0,0)\begin{pmatrix} 2.26 & -2.26 & -2.26 & -2.26 & -2.26 \\ & 4.53 & 2.26 & 2.26 & 2.26 \\ & & 4.53 & 2.26 & 2.26 \\ & & & 4.53 & 2.26 \\ & & & & 4.53 \end{pmatrix}\begin{pmatrix}1\\1\\0\\0\\0\end{pmatrix} \\
&= (0, 2.26, 0, 0, 0)\begin{pmatrix}1\\1\\0\\0\\0\end{pmatrix} = \\
&= 2.26 \\
\widehat{SE}(\widehat{\beta}_0 + \widehat{\beta}_1) &= \sqrt{2.26} = 1.5 = \sqrt{MS(E)/4}
\end{aligned}
$$

(Same for all sample treatment means in balanced experiment.)

```
proc glm data=one;
   class drug;
   model  y=drug/solution inv;
run;
```

```
The SAS System
The GLM Procedure

     Class Level Information
Class          Levels    Values
drug                5    1 2 3 4 5

                                         Sum of
Source                      DF          Squares     Mean Square    F Value    Pr > F
Model                        4      1480.823000      370.205750      40.88    <.0001
Error                       15       135.822500        9.054833
Corrected Total             19      1616.645500

R-Square     Coeff Var      Root MSE        y Mean
0.915985      13.12023      3.009125      22.93500

Source                      DF        Type I SS     Mean Square    F Value    Pr > F
drug                         4      1480.823000      370.205750      40.88    <.0001

                                       Standard
Parameter           Estimate              Error    t Value    Pr > |t|
Intercept        27.80000000 B       1.50456251      18.48      <.0001
drug      1       0.80000000 B       2.12777270       0.38      0.7122
drug      2       3.57500000 B       2.12777270       1.68      0.1136
drug      3     -19.97500000 B       2.12777270      -9.39      <.0001
drug      4      -8.72500000 B       2.12777270      -4.10      0.0009
drug      5       0.00000000 B        .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to
      solve the normal equations.  Terms whose estimates are followed by the letter 'B'
      are not uniquely estimable.
```

$\lambda_1 = (1,1,0,0,0,0)$ $\Longrightarrow \lambda_1' \beta = \mu + \tau_1$ is estimable

$\lambda_2 = (0,-1,1,0,0,0)$

$= X_{5.} - X_{4.}$

$\lambda_2' \beta$ estimable

$\lambda_3 = (0,1,0,0,0,0)$

$\lambda_3' \beta$ nonestimable

$= \tau_1$

$$y = \begin{pmatrix} 29.6 \\ 24.3 \\ 28.5 \\ 32.0 \\ 27.3 \\ 32.6 \\ 30.8 \\ 34.8 \\ 5.8 \\ 6.2 \\ 11.0 \\ 8.3 \\ 21.6 \\ 17.4 \\ 18.3 \\ 19.0 \\ 29.2 \\ 32.8 \\ 25.0 \\ 24.2 \end{pmatrix}, X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \beta = \begin{pmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \end{pmatrix},$$

$$X'X = \begin{pmatrix} 20 & 4 & 4 & 4 & 4 & 4 \\ 4 & 4 & 0 & 0 & 0 & 0 \\ 4 & 0 & 4 & 0 & 0 & 0 \\ 4 & 0 & 0 & 4 & 0 & 0 \\ 4 & 0 & 0 & 0 & 4 & 0 \\ 4 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}$$

Add rows 2-6 of $X'X$. $(X'X)^{-1}$?

```
NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to
      solve the normal equations.  Terms whose estimates are followed by the letter 'B'
      are not uniquely estimable.
```

```
     Class Level Information
Class          Levels     Values
drug                5     1 2 3 4 5
```

X'X Generalized Inverse (g2)

	Intercept	drug 1	drug 2	drug 3	drug 4	drug 5	y
Intercept	0.25	-0.25	-0.25	-0.25	-0.25	0	27.8
drug 1	-0.25	0.5	0.25	0.25	0.25	0	0.8
drug 2	-0.25	0.25	0.5	0.25	0.25	0	3.575
drug 3	-0.25	0.25	0.25	0.5	0.25	0	-19.975
drug 4	-0.25	0.25	0.25	0.25	0.5	0	-8.725
drug 5	0	0	0	0	0	0	0

A generalized inverse, of a matrix A, A^- has this property: $AA^-A = A$.

$$(X'X)^- X'Y = \begin{pmatrix} 0.25 & -0.25 & -0.25 & -0.25 & -0.25 & 0 \\ -0.25 & 0.5 & 0.25 & 0.25 & 0.25 & 0 \\ -0.25 & 0.25 & 0.5 & 0.25 & 0.25 & 0 \\ -0.25 & 0.25 & 0.25 & 0.5 & 0.25 & 0 \\ -0.25 & 0.25 & 0.25 & 0.25 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 458.7 \\ 114.4 \\ 125.5 \\ 31.3 \\ 76.3 \\ 111.2 \end{pmatrix} = \begin{pmatrix} 27.800 \\ 0.800 \\ 3.575 \\ -19.975 \\ -8.725 \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} \widehat{\mu + \tau_5} \\ \widehat{\tau_1 - \tau_5} \\ \widehat{\tau_2 - \tau_5} \\ \widehat{\tau_3 - \tau_5} \\ \widehat{\tau_4 - \tau_5} \end{pmatrix}$$

The $\widehat{\tau}_i$ are not uniquely estimable, and will change with different choices for the generalized inverse, $(X'X)^-$, but $\widehat{\mu + \tau_i}$ (and many other functions of interest), are uniquely estimable, and will not change with different generalized inverses.

Generalized inverse corresponding to dropped 5^{th} row:

```
The REG Procedure
Model: MODEL1
                         Model Crossproducts X'X X'Y Y'Y

Variable       Intercept          x1          x2          x3          x4           y

Intercept             20           4           4           4           4       458.7
x1                     4           4           0           0           0       114.4
x2                     4           0           4           0           0       125.5
x3                     4           0           0           4           0        31.3
x4                     4           0           0           0           4        76.3
y                  458.7       114.4       125.5        31.3        76.3    12136.93

                        Parameter Estimates

                       Parameter     Standard
Variable     DF         Estimate        Error    t Value    Pr > |t|

Intercept     1         27.80000      1.50456      18.48      <.0001
x1            1          0.80000      2.12777       0.38      0.7122
x2            1          3.57500      2.12777       1.68      0.1136
x3            1        -19.97500      2.12777      -9.39      <.0001
x4            1         -8.72500      2.12777      -4.10      0.0009
```

Generalized inverse corresponding to dropped 1^{st} row:

```
The REG Procedure
Model: MODEL2
                         Model Crossproducts X'X X'Y Y'Y

Variable        Intercept            x2            x3            x4            x5             y

Intercept              20             4             4             4             4         458.7
x2                      4             4             0             0             0         125.5
x3                      4             0             4             0             0          31.3
x4                      4             0             0             4             0          76.3
x5                      4             0             0             0             4         111.2
y                   458.7         125.5          31.3          76.3         111.2      12136.93

                   Parameter Estimates

                      Parameter       Standard
Variable      DF       Estimate          Error    t Value    Pr > |t|

Intercept      1       28.60000        1.50456      19.01      <.0001
x2             1        2.77500        2.12777       1.30      0.2118
x3             1      -20.77500        2.12777      -9.76      <.0001
x4             1       -9.52500        2.12777      -4.48      0.0004
x5             1       -0.80000        2.12777      -0.38      0.7122
```

Generalized inverse corresponding to dropped 2^{nd} row:

```
The REG Procedure
Model: MODEL3
                         Model Crossproducts X'X X'Y Y'Y

Variable        Intercept              x1              x3              x4              x5               y

Intercept              20               4               4               4               4           458.7
x1                      4               4               0               0               0           114.4
x3                      4               0               4               0               0            31.3
x4                      4               0               0               4               0            76.3
x5                      4               0               0               0               4           111.2
y                   458.7           114.4            31.3            76.3           111.2        12136.93

                        Parameter Estimates

                     Parameter        Standard
Variable      DF      Estimate           Error    t Value    Pr > |t|

Intercept      1      31.37500         1.50456      20.85      <.0001
x1             1      -2.77500         2.12777      -1.30      0.2118
x3             1     -23.55000         2.12777     -11.07      <.0001
x4             1     -12.30000         2.12777      -5.78      <.0001
x5             1      -3.57500         2.12777      -1.68      0.1136
```

Generalized inverse corresponding to dropped 3^{rd} row:

```
The REG Procedure
Model: MODEL4
                              Model Crossproducts X'X X'Y Y'Y

Variable        Intercept             x1             x2             x4             x5              y

Intercept              20              4              4              4              4          458.7
x1                      4              4              0              0              0          114.4
x2                      4              0              4              0              0          125.5
x4                      4              0              0              4              0           76.3
x5                      4              0              0              0              4          111.2
y                   458.7          114.4          125.5           76.3          111.2       12136.93

                        Parameter Estimates

                       Parameter       Standard
Variable      DF        Estimate          Error    t Value    Pr > |t|

Intercept      1         7.82500        1.50456       5.20      0.0001
x1             1        20.77500        2.12777       9.76      <.0001
x2             1        23.55000        2.12777      11.07      <.0001
x4             1        11.25000        2.12777       5.29      <.0001
x5             1        19.97500        2.12777       9.39      <.0001
```

Generalized inverse corresponding to dropped 4^{th} row:

```
The REG Procedure
Model: MODEL5
                              Model Crossproducts X'X X'Y Y'Y

Variable        Intercept              x1              x2              x3              x5               y

Intercept              20               4               4               4               4           458.7
x1                      4               4               0               0               0           114.4
x2                      4               0               4               0               0           125.5
x3                      4               0               0               4               0            31.3
x5                      4               0               0               0               4           111.2
y                   458.7           114.4           125.5            31.3           111.2        12136.93

                        Parameter Estimates

                      Parameter       Standard
Variable      DF       Estimate          Error    t Value    Pr > |t|

Intercept      1       19.07500        1.50456      12.68      <.0001
x1             1        9.52500        2.12777       4.48      0.0004
x2             1       12.30000        2.12777       5.78      <.0001
x3             1      -11.25000        2.12777      -5.29      <.0001
x5             1        8.72500        2.12777       4.10      0.0009
```

Generalized inverse corresponding to dropped intercept

The REG Procedure
Model: MODEL6

Model Crossproducts X'X X'Y Y'Y

Variable	x1	x2	x3	x4	x5	y
x1	4	0	0	0	0	114.4
x2	0	4	0	0	0	125.5
x3	0	0	4	0	0	31.3
x4	0	0	0	4	0	76.3
x5	0	0	0	0	4	111.2
y	114.4	125.5	31.3	76.3	111.2	12136.93

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
x1	1	28.60000	1.50456	19.01	<.0001
x2	1	31.37500	1.50456	20.85	<.0001
x3	1	7.82500	1.50456	5.20	0.0001
x4	1	19.07500	1.50456	12.68	<.0001
x5	1	27.80000	1.50456	18.48	<.0001

As an exercise, obtain the least squares estimate of $\mu + \tau_1$ and $\theta_2 = \tau_2 - \tau_1$ using each generalized inverse:

Gen'd Inverse	$\widehat{\mu + \tau_1}$	$\widehat{\tau_2 - \tau_1}$
1	$27.8 + 0.8 = 28.6$	$3.575 - 0.8 = 2.775$
2	$28.6 + 0 = 28.6$	$2.775 - 0.0 = 2.775$
3	$31.375 - 2.775 = 28.6$	$0 - (-2.775) = 2.775$
4		
5		
6		

Apparently, $\mu, \tau_1, \ldots$ are not uniquely estimable, but $\mu + \tau_1$ and $\tau_2 - \tau_1$ are.

Complete this table as an exercise.

Estimable functions of regression coefficients

$E(Y_{ij}) = \mu + \tau_i$ $\qquad E(Y) = X\beta$ $\qquad \beta = \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_5 \end{pmatrix}$

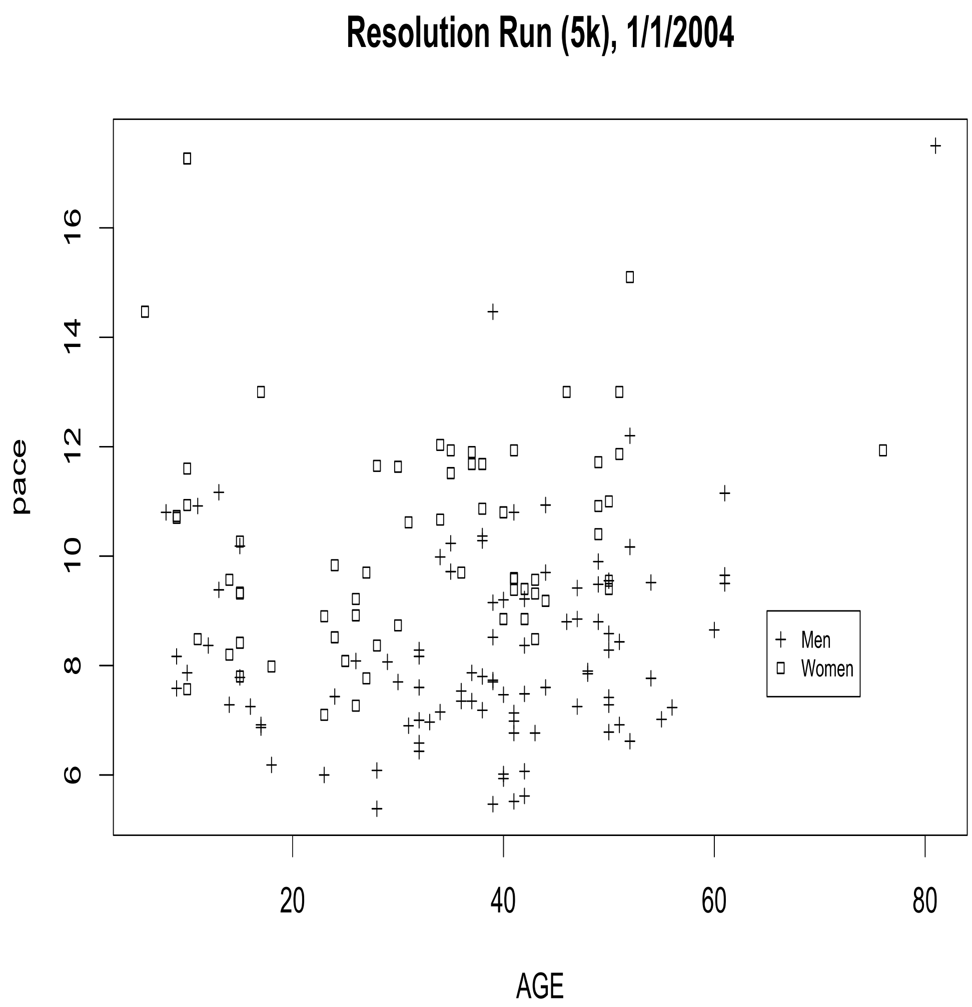
- Definition: In the general linear model, $Y = X\beta + E$, a linear combination of parameters, $\lambda'\beta$ is said to be estimable if there exists a linear combination of the data, $a'Y$ with that as its expectation, so that $E(a'Y) = \lambda'\beta$. If no such linear combination exists, $\lambda'\beta$ is nonestimable.
- If $\lambda \in \text{rowsp}(X)$ then $\lambda'\beta$ is estimable. (λ can be obtained as a linear combination of the rows of X.)

not uniquely estimable

1. Is $\lambda'\beta$ estimable where $\lambda' = (1, 1, 0, 0, 0, 0)$ yes Which rows of X? 1st
2. Is $\lambda'\beta$ estimable where $\lambda' = (0, -1, 1, 0, 0, 0)$ yes Which rows of X? 5th, 1th
3. Is $\lambda'\beta$ estimable where $\lambda' = (0, 1, 0, 0, 0, 0)$ No Which rows of X?

A general linear model

Both indicator variables for factorial effects (sex) as well as continously valued variables (age, age squared).



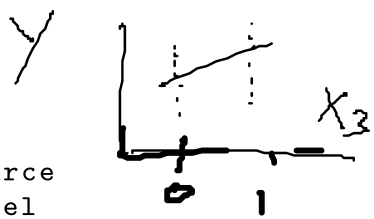
Quadratic model $\mu(x) = \beta_0 + \beta_1 x + \beta_1 x^2$ used for association between mean pace and age. How could the model be extended to incorporate sex differences? Let $x_2 = x^2$ and let an indicator variable x_3 be defined by

$$x_3 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

Some candidate models:

$$\begin{aligned}
\mu(x_1, x_2, x_3) &= \beta_0 \\
\mu(x_1, x_2, x_3) &= \beta_0 + \beta_3 x_3 \\
\mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 \\
\mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\
\mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad \hat{\mu} = 10.2 + \\
\mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \underbrace{x_1 x_3}_{x_4} + \beta_5 \underbrace{x_2 x_3}_{x_5}
\end{aligned}$$

```
data race5k; set race5k;
   sexf=(sex="F"); age2=age*age;  agef=age*sexf;  age2f=age2*sexf;
proc reg data=one ;
   model pace=;
   model pace=sexf;  /* equivalent to two-sample t-test */
   model pace=age age2;
   model pace=sexf age age2;
   model pace=sexf age age2 agef age2f;
   test agef=0, age2f=0;
run;
```




```
                        The REG Procedure
                         Model: MODEL1

                                 Sum of           Mean
Source                 DF       Squares         Square    F Value    Pr > F
Model                   0             0              .          .         .
Error                 159     788.09472        4.95657
Corrected Total       159     788.09472

                       Parameter       Standard
  Variable      DF      Estimate          Error    t Value    Pr > |t|
  Intercept      1       9.12063        0.17601      51.82      <.0001

                         Model: MODEL2

                                 Sum of           Mean
Source                 DF       Squares         Square    F Value    Pr > F
Model                   1     170.74137      170.74137      43.70    <.0001
Error                 158     617.35335        3.90730
Corrected Total       159     788.09472

          Root MSE              1.97669    R-Square     0.2167

                       Parameter       Standard
  Variable      DF      Estimate          Error    t Value    Pr > |t|
  Intercept      1       8.26614        0.20280      40.76      <.0001
  sexf           1       2.10335        0.31819       6.61      <.0001
```

$$t = \frac{\hat{Y}_F - \hat{Y}_M}{\sqrt{S_p^2\left(\frac{1}{n_F}+\frac{1}{n_M}\right)}}$$

$$\left(S_p^2 = \frac{1}{n_F-1}S_F^2 + \frac{1}{n_M-1}S_M^2\right)$$

$$t = \frac{2.1}{.32} = 6.61$$

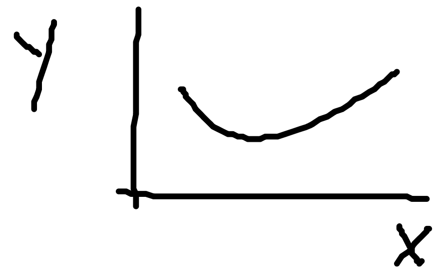
Model: MODEL4

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	113.64500	56.82250	13.23	<.0001
Error	157	674.44972	4.29586		
Corrected Total	159	788.09472			

Root MSE	2.07265	R-Square	0.1442

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	11.78503	0.70216	16.78	<.0001
age	1	-0.19699	0.04113	-4.79	<.0001
age2	1	0.00294	0.00057380	5.12	<.0001

parabola



Model: MODEL5

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	290.34851	96.78284	30.33	<.0001
Error	156	497.74621	3.19068		
Corrected Total	159	788.09472			

Root MSE	1.78625	R-Square	0.3684

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.18317	0.64228	15.85	<.0001
sexf	1	2.19792	0.29535	7.44	<.0001
age	1	-0.17146	0.03562	-4.81	<.0001
age2	1	0.00281	0.00049481	5.67	<.0001

parallel parabolas

$F = \frac{3.2/2}{3.2} = .5$

$$\mu(x) = \begin{cases} 10.18 - 0.17x + 0.0028x^2 & \text{for men} \\ \dfrac{10.18 + 2.2(1) - .17x + .0028x^2}{12.4 - .17x + .0028x^2} & \text{for women} \end{cases}$$

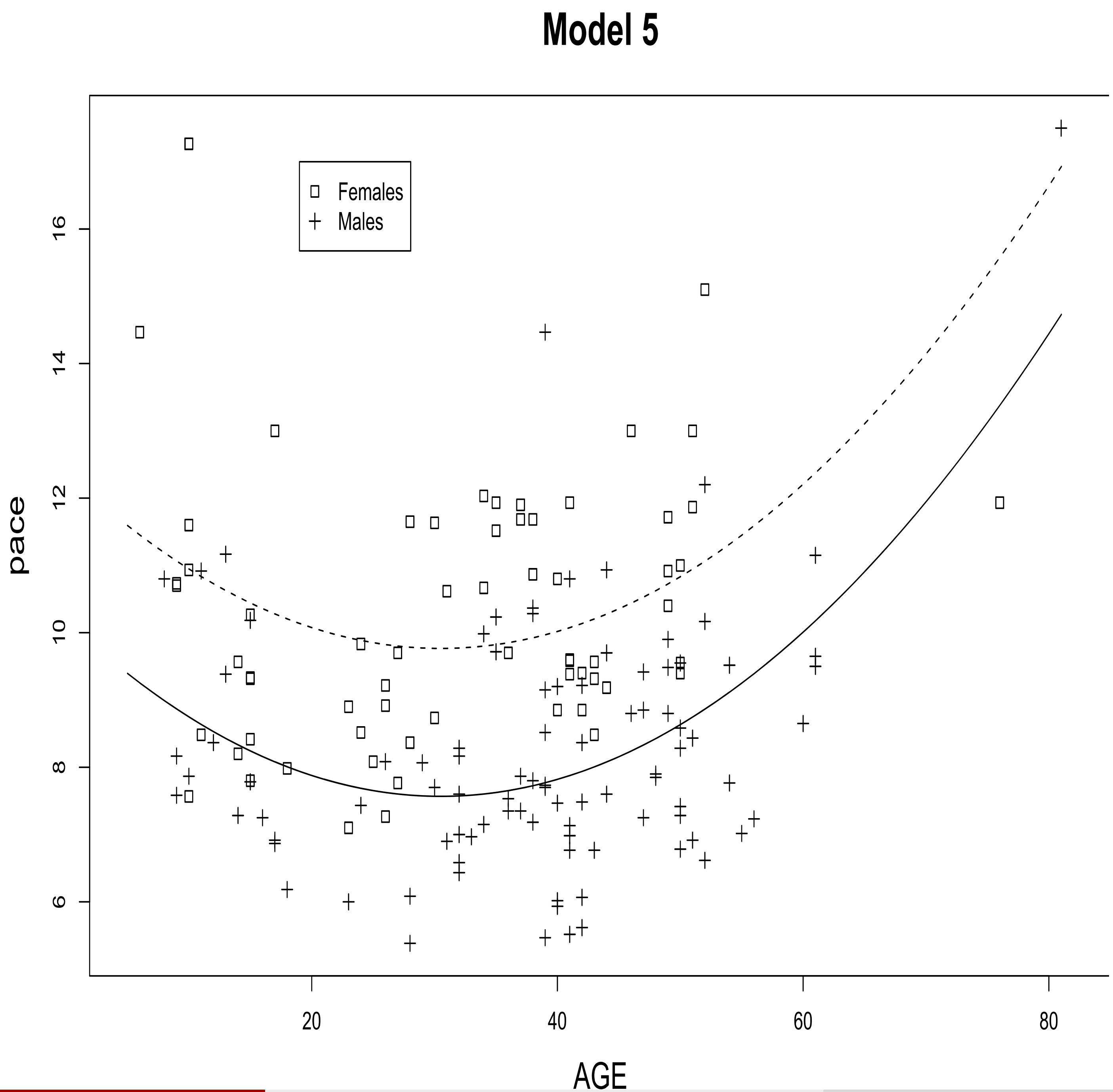
Model: MODEL6

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	293.52828	58.70566	18.28	<.0001
Error	154	494.56644	3.21147		
Corrected Total	159	788.09472			

Root MSE	1.79206	R-Square	0.3725

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.60848	0.88641	11.97	<.0001
sexf	1	1.25728	1.23237	1.02	0.3092
age	1	-0.19986	0.04842	-4.13	<.0001
age2	1	0.00321	0.00064628	4.96	<.0001
agef	1	0.06882	0.07298	0.94	0.3471
age2f	1	-0.00103	0.00103	-0.99	0.3217

$$\mu(x) = \begin{cases} \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3(0) + \beta_4(0) + \beta_5(0) & \text{men} \\ \boxed{10.61 - 0.20x + 0.0032x^2} & \\ \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3(1) + \beta_4(x) + \beta_5(x^2)(1) & \text{women} \\ \beta_0 + \beta_3 + (\beta_1 + \beta_4)x + (\beta_2 + \beta_5)x^2 & \\ 10.61 + 1.25 + (-0.20 + 0.07)x + (0.0032 - 0.0010)x^2 & \\ \boxed{11.86 - 0.13x + 0.0022x^2} & \end{cases}$$



Model 5
Females
Males
pace
6
8
10
12
14
16
20
40
60
80
AGE

Comparison of models 5 and 6

parallel

$$
\begin{aligned}
\text{reduced: } \mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\
\text{full: } \mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3
\end{aligned}
$$

$H_0: \beta_4 = \beta_5 = 0$

Extra sum of squares:

$$R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2, \beta_3) = SS[R]_f - SS[R]_r = 293.5 - 290.3 = 3.0$$

The F-ratio

$$F = \frac{R(\beta_4, \beta_5 | \beta_0, \beta_1, \beta_2, \beta_3)/(5-3)}{MS[E]_f} = \frac{3.2/2}{3.21} = \frac{1.6}{3.21} = 0.5$$

The observed F-ratio is not significant on $df = 2, 154$.

In SAS, you could use

```
proc reg;
   model pace=age age2 sexf agef age2f;
   test agef=0, age2f=0;
run;
```

to get the following model selection F-ratio in the output:

```
                         The REG Procedure
                          Model: MODEL6

              Test 1 Results for Dependent Variable pace
```

Nonlinear functions of parameters

Estimate "peak" running age for men/women. θ_M and θ_W denote peak running ages for men and women respectively. Using calculus on the model 6 regression,

$$\hat{\mu}_M(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \qquad \hat{\mu}_M'(x) = \hat{\beta}_1 + 2\hat{\beta}_2 x = 0$$

$$\Rightarrow x = -\hat{\beta}_1 / 2\hat{\beta}_2 = \frac{-(-2)}{2(.0032)}$$

$$\hat{\mu}_F(x) = (\hat{\beta}_0 + \hat{\beta}_3) + (\hat{\beta}_1 + \hat{\beta}_4) x + (\hat{\beta}_2 + \hat{\beta}_5) x^2$$

$$\hat{\mu}_F'(x) = 0 \Rightarrow x = \frac{-(\hat{\beta}_1 + \hat{\beta}_4)}{2(\hat{\beta}_2 + \hat{\beta}_5)} = \frac{-(-13)}{2(.0022)} = 30.1$$

reduced model $\hat{\mu}'(x) = 0 \Rightarrow x = \frac{-(.17)}{2(.0028)} = 30.5$

$$\theta_M = \underline{\hspace{4cm}}, \quad \theta_W = \underline{\hspace{8cm}}$$

These are nonlinear functions of regression parameters. Note that acceptance of any model but 6 implies equality of these peak ages.

$$\hat{\theta}_W = \begin{cases} 30.5 & \text{different intercepts model (5)} \\ 30.1 & \text{full model (6)} \end{cases}$$

$$\hat{\theta}_M = \begin{cases} 30.5 & \text{different intercepts model (5)} \\ 31.1 & \text{full model (6)} \end{cases}$$

LOF

Lack-of-fit of a polynomial regression model

Completely randomized experiment in White Mountains of NH. $n = 20$ lanes of dimension $0.5m \times 1.5m$ randomized to 5 foot-traffic treatments:

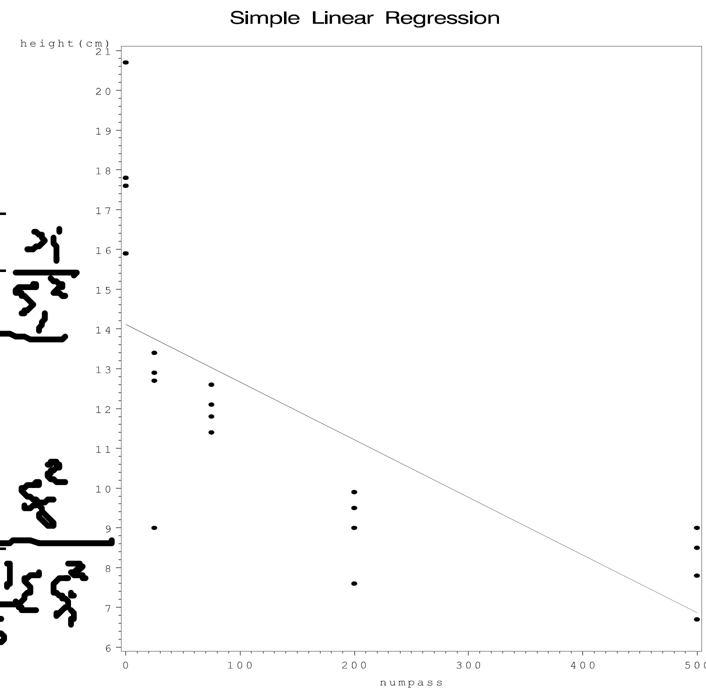
- $E(Y_{ij}) = \mu + \tau_i$
- $E(Y_{ij}) = \beta_0 + \beta_1 x_i$

i : trt grp	x : # of passes	y_{ij} : Height(cm)			
1	0	20.7	15.9	17.8	17.6
2	25	12.9	13.4	12.7	9.0
3	75	11.8	12.6	11.4	12.1
4	200	7.6	9.5	9.9	9.0
5	500	7.8	9.0	8.5	6.7

$\bar{y}_i$, s_i^2

s_i^2

$MS(E) = \frac{1}{5}\sum s_i^2$



Two models for mean plant height:

$df_e = 18$

$df_e = 15$

$$\text{SLR model} : \mu(x) = \beta_0 + \beta_1 x_i$$

$$\text{one-factor ANOVA model} : \mu_{ij} = \mu + \tau_i$$

$\Delta df = 3$

most complex possible

```
proc reg data=one;
   model y=numpass;

proc glm data=one;              *find the F-ratio that compares SLR with one-factor model;
   class cnumpass;              *find SS(Trt);
   model y=numpass cnumpass;            run;
```

```
                        Analysis of Variance

                                  Sum of           Mean
Source                   DF      Squares         Square    F Value    Pr > F
Model                     1    141.29532      141.29532      19.15    0.0004
Error                    18    132.79418        7.37745
Corrected Total          19    274.08950

                                 Parameter      Standard
 Variable    Label         DF     Estimate         Error    t Value    Pr > |t|
 Intercept   Intercept      1     14.11334       0.80592      17.51      <.0001
 numpass                    1     -0.01449       0.00331      -4.38      0.0004

                        The GLM Procedure

                   Class         Levels    Values
                   cnumpass           5    0 25 75 200 500
                                    Sum of
Source                      DF     Squares     Mean Square   F Value   Pr > F
Model                        4   243.1620000    60.7905000     29.48   <.0001
Error                       15    30.9275000     2.0618333
Corrected Total             19   274.0895000

Source                      DF     Type I SS   Mean Square   F Value   Pr > F
numpass                      1   141.2953228   141.2953228     68.53   <.0001
cnumpass                     3   101.8666772    33.9555591     16.47   <.0001

Source                      DF   Type III SS   Mean Square   F Value   Pr > F
numpass                      0     0.0000000         .          .        .
cnumpass                     3   101.8666772    33.9555591     16.47   <.0001
```

$$F = \frac{(243.2-141.3)/3}{2.06} \doteq 16.47 \doteq \frac{101.9/3}{2.06}$$

SS(Pure error)

SS(LOF)

$df = 3, 15$

LOF!!

When t treatments have interval scale, the SLR model, and all polynomials of degree $p \leq t-2$, are nested in one-factor ANOVA model with t treatment means.

<u>F-ratio for lack-of-fit</u>

To test for lack-of-fit of a polynomial (*reduced*) model of degree p, use extra sum-of-squares F-ratio on $t-1-p$ and $N-t$ df:

exercise

$$F = \frac{SS[\text{lack of fit}]/(t-1-p)}{MS[\text{pure error}]}, \quad \text{where}$$

test $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ for LOF

$$MS[\text{pure error}] = MS[E]_{full} \quad \text{and}$$

on

$$\begin{aligned} SS[\text{lack-of-fit}] &= SS[Trt] - SS[R]_{poly} = SS[E]_{poly} - SS[E]_{full} \\ &= SS[E]_{poly} - SS[\text{pure error}] \end{aligned}$$

$dF = 2, 15$

In a simple linear ($p = 1$) model for the meadows data,

$$SS[\text{lack of fit}] = 243.163 - 141.295 = 101.867 \quad \text{on} \quad t-1-p = 3df$$

and the sum of squares for pure error is $SS[E]_{full} = 30.93$ yielding

$$F = \frac{101.867/3}{30.93/15} \approx \frac{34}{2.1} = 16.5.$$

(highly significant since $F(0.01, 3, 15) = 5.42$.)

$\boxed{\Longrightarrow \text{model misspecified: SLR model suffers from lack of fit.}}$

Some terminology for factorial experiments:

- contrasts
- orthogonal contrasts
- multiple contrasts
- expected mean squares
- familywise or experimentwise error rates
- power

Comparisons (contrasts) among means

Definition: In the one-way ANOVA layout:

$$Y_{ij} = \mu_i + E_{ij}, i = 1, 2, \ldots, t, \text{ and } j = 1, 2, \ldots, n_i$$

with $E_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, A linear combination of treatment means, $\theta = \sum c_i \mu_i$ is called a Contrast if .

$$c_1 + c_2 + \cdots + c_t = \sum_1^t c_j = 0.$$

Definition: Contrasts in which only two of the coefficients are nonzero are called simple or Pairwise contrasts. Example: $\mu_1 - \mu_2$

Definition: Contrasts in with more than two nonzero coefficients are called complex contrasts. Example: $\mu_1 - \mu_2 - (\mu_3 - \mu_4)$

Result: The *best* estimator for a contrast of interest obtained by substituting treatment group sample means $\bar{y}_{i+}$ for treatment population means μ_i in the contrast θ:

$$\hat{\theta} = c_1 \bar{Y}_{1+} + c_2 \bar{Y}_{2+} + \cdots + c_t \bar{Y}_{t+}.$$

For binding fractions, contrast penicillin and Tetracyclin (population) means

$$\theta = \mu_1 - \mu_2 = (1)\mu_1 + (-1)\mu_2 + (0)\mu_3 + (0)\mu_4 + (0)\mu_5$$

Using the result, point estimator of θ is

$$\hat{\theta} = \hat{\mu}_1 - \hat{\mu}_2 = \bar{Y}_{1+} - \bar{Y}_{2+}$$

Antibiotic	Binding Percentage				Sample mean	Sample variance
Penicillin G	29.6	24.3	28.5	32	28.6	10.4
Tetracyclin	27.3	32.6	30.8	34.8	31.4	10.1
Streptomycin	5.8	6.2	11	8.3	7.8	5.7
Erythromycin	21.6	17.4	18.3	19	19.1	3.3
Chloramphenicol	29.2	32.8	25	24.2	27.8	15.9

Source	d.f.	Sum of squares	Mean Square	F
Treatments	4	1481	370	41
Error	15	136	9.05	
Total	19	1617		

Substitution of $\bar{y}_{1+}$ and $\bar{y}_{2+}$ yields $\hat{\theta} = 28.6 - 31.4 = -2.8$.
Q: How *good* is this estimate? (Quantify the associated uncertainty.)

Sampling distribution of $\hat{\theta}$

$\widehat{\theta}$ a linear combo of independent averages of normals, hence normal with std.err.

$$SE(\hat{\theta}) = \sqrt{\frac{c_1^2}{n_1}\sigma^2 + \frac{c_2^2}{n_2}\sigma^2 + \cdots + \frac{c_t^2}{n_t}\sigma^2} = \sqrt{\sigma^2 \sum_{i=1}^{i=t} \frac{c_i^2}{n_i}},$$

estimated by

$$\hat{SE}(\hat{\theta}) = \sqrt{MS[E] \sum_{i=1}^{i=t} \frac{c_i^2}{n_i}}$$

To test $H_0 : \theta = \theta_0$ (often 0) versus $H_1 : \theta \neq \theta_0$ a t use t-test:

$$t = \frac{\text{est} - \text{null}}{\hat{SE}} = \frac{\hat{\theta} - \theta_0}{\hat{SE}(\hat{\theta})} \overset{H_0}{\sim} t_{N-t}.$$

At level α, the critical value for this test is $t(N - t, \alpha/2)$.

$100(1 - \alpha)\%$ confidence interval for a contrast $\theta = \sum c_i \mu_i$ given by

$$\boxed{\pm t(\alpha/2, N - t)\sqrt{MS(E)}}$$

Here,

$$\widehat{SE}(\hat{\theta}) = \sqrt{\left(\frac{1^2}{n_1} + \frac{(-1)^2}{n_2}\right)(9.05)} = \sqrt{\frac{9.05}{2}} = 2.127$$

So that the t statistic becomes

$$\frac{-2.8}{2.127} = -1.32$$

which is not in the critical region, so that the sample mean binding fractions for Penicillin G and Tetracyclin do not differ significantly.

A 95% confidence interval is given by

$$-2.8 \pm 2.13(2.127) \text{ or } (-7.3, 1.7)$$

Code (next page) estimates all $\boxed{\text{pairwise}}$ contrasts involving Pen. G:

- $\theta_1 = a'\mu = (1, -1, 0, 0, 0)\mu$
- $\theta_2 = b'\mu = ?$
- $\theta_3 = c'\mu = ?$
- $\theta_4 = d'\mu = ?$

along with complex contrast comparing Pen G. with mean of other four antibiotics:

$$\theta_5 = (\quad , \quad , \quad , \quad , \quad)\mu$$

Here $\mu = (\mu_1, \mu_2, \mu_3, \mu_4, \mu_5)'$.

```
proc glm data=one;
   class drug;
   model y=drug/clparm;
   estimate "theta1" drug 1 -1;
   estimate "theta2" drug 1 0 -1;
   estimate "theta3" drug 1 0 0 -1;
   estimate "theta4" drug 1 0 0 0 -1;
   estimate "theta5" drug 4 -1 -1 -1 -1/divisor=4;
run;
```

```
The GLM Procedure

     Class Level Information

Class          Levels    Values
drug                5    1 2 3 4 5

                                       Sum of
Source                    DF          Squares     Mean Square    F Value    Pr > F
Model                      4      1480.823000      370.205750      40.88    <.0001
Error                     15       135.822500        9.054833
Corrected Total           19      1616.645500

R-Square     Coeff Var      Root MSE        y Mean
0.915985      13.12023      3.009125      22.93500


Source                    DF        Type I SS     Mean Square    F Value    Pr > F
drug                       4      1480.823000      370.205750      40.88    <.0001

                                           Standard
Parameter                 Estimate            Error    t Value    Pr > |t|      95% Confidence Limits
theta1                  -2.7750000       2.12777270      -1.30      0.2118      -7.3102402     1.7602402
theta2                  20.7750000       2.12777270       9.76      <.0001      16.2397598    25.3102402
theta3                   9.5250000       2.12777270       4.48      0.0004       4.9897598    14.0602402
theta4                   0.8000000       2.12777270       0.38      0.7122      -3.7352402     5.3352402
theta5                   7.0812500       1.68215202       4.21      0.0008       3.4958278    10.6666722
```

Orthogonal contrasts: Let two contrasts θ_1 and θ_2 be given by

$$\theta_1 = c_1\mu_1 + \cdots + c_t\mu_t \qquad \text{and} \qquad \theta_2 = d_1\mu_1 + \cdots + d_t\mu_t$$

Definition: The two contrasts θ_1 and θ_2 are mutually orthogonal if the products of their coefficients sum to zero: $c_1 d_1 + \cdots c_t d_t = \sum_{i=1}^t c_i d_i = 0$. A *set* of several contrasts $\theta_1, \ldots, \theta_k$ is mutually orthogonal if all pairs mutually orthogonal.

$$(-1, 1, 0, 0, 0) \text{ and } (0, 0, -1, 1, 0) \text{ orthogonal ?}$$

$$(1, -1/2, -1/2, 0, 0) \text{ and } (0, 0, 0, -1, 1) \text{ orthogonal ?}$$

$$(-1, 1, 0, 0, 0) \text{ and } (0, -1, 1, 0, 0) \text{ orthogonal ?}$$

θ_i and θ_j orthogonal $\Longrightarrow$ $\hat{\theta}_i$ and $\hat{\theta}_j$ are statistically independent.

Contrast sums of squares

As $SS[Trt]$ quantifies treatment effect, $SS(\theta_i)$ quantifies contrast effect:

$$SS[\hat{\theta}_1] = \frac{\hat{\theta}_1^2}{\left(\frac{c_1^2}{n_1} + \cdots + \frac{c_t^2}{n_t}\right)}$$

If $\theta_1, \ldots, \theta_{t-1}$ are $t-1$ mutually orthogonal contrasts, then

$$SS[Trt] = SS(\hat{\theta}_1) + SS(\hat{\theta}_2) + \cdots + SS(\hat{\theta}_{t-1})$$

For single df contrasts, if $H_0 : \theta_i = 0$,

$$E(SS[\hat{\theta}_j]) = \sigma^2.$$

To test $H_0 : \theta_j = 0$ versus $H_1 : \theta_j \neq 0$, use F below, with $df =$ ________, ________:

$$F = \frac{SS[\hat{\theta}_j]}{MS[E]}$$

For $\theta_1 = \mu_1 - \mu_2$ in the binding fractions,

$$F = \frac{(-2.8)^2}{MS[E]\left(\frac{1}{4} + \frac{(-1)^2}{4} + 0 + 0 + 0\right)} = 1.73.$$

(Using $F(0.05, 1, 15) = 4.54$, is $H_0 : \theta_1 = 0$ plausible?)

Number of contaminants in IV fluids made by $t = 3$ pharmaceutical companies

	Cutter	Abbott	McGaw
	255	105	577
	264	288	515
	342	98	214
	331	275	413
	234	221	401
	217	240	260
$\bar{y}_{i+}$	273.8	204.5	396.7

Source	d.f.	Sum of squares	Mean Square	F
Treatments (or pharmacies)	2	113646	56823	
Error	15	146753	9784	
Total	17	260400		

Consider the following 2 contrasts:

$$\theta_1 = \mu_M - \mu_A \quad \text{and} \quad \theta_2 = \mu_C - \frac{\mu_M + \mu_A}{2}$$

Q: Are these contrasts orthogonal?

Q: Are the estimated contrasts $\hat{\theta}_1$ and $\hat{\theta}_2$ independent?

Exercise: Compute $SS[\hat{\theta}_1]$ and $SS[\hat{\theta}_2]$. Add em up.

```
proc glm order=formatted;
   title "contaminant particles in IV fluids";
   class firm;
   model con=firm;
   contrast 'C - avg of M and A' firm -0.5 1 -0.5;
   contrast 'McGaw - Abbott' firm -1 0 1;
   estimate 'C - avg of M and A' firm -0.5 1 -0.5;
   estimate 'McGaw - Abbott' firm -1 0 1;
run;
```

```
                    contaminant particles in IV fluids                   1
                             The GLM Procedure

               Class          Levels     Values
               firm                3     Abbott Cutter McGaw

                                     Sum of
 Source                   DF        Squares    Mean Square  F Value  Pr > F
 Model                     2    113646.3333     56823.1667     5.81  0.0136
 Error                    15    146753.6667      9783.5778
 Corrected Total          17    260400.0000

            R-Square      Coeff Var      Root MSE      con Mean
            0.436430      33.91268       98.91197      291.6667

 Contrast                 DF    Contrast SS    Mean Square  F Value  Pr > F
 C - avg of M and A        1      2862.2500      2862.2500     0.29  0.5965
 McGaw - Abbott            1    110784.0833    110784.0833    11.32  0.0043

                                            Standard
Parameter                   Estimate           Error    t Value    Pr > |t|
C - avg of M and A        -26.750000      49.4559849      -0.54      0.5965
McGaw - Abbott            192.166667      57.1068524       3.37      0.0043
```

Multiple Comparisons

- Too many tests of significance brings creeping type I error rate
- e.g. consider the case with $t = 5$ (antibiotic treatments): all simple (pairwise) contrasts of the form $\theta = \mu_i - \mu_j$
- $\begin{pmatrix} 5 \\ 2 \end{pmatrix} =$ ________ tests of significance each at level $\alpha = 0.05$

When testing k contrasts, the $\boxed{\text{experimentwise error rate}}$ (or familywise) is

$$fwe = \Pr(\qquad\qquad\qquad\qquad\qquad\qquad)$$

Methods for simultaneous inference for multiple contrasts include

- Bonferroni
- Tukey
- Scheffé (won't cover)

When the number of comparisons is in the thousands, and FWE control is hopeless, more manageable type I error rate is the False Discovery Rate (**FDR**):

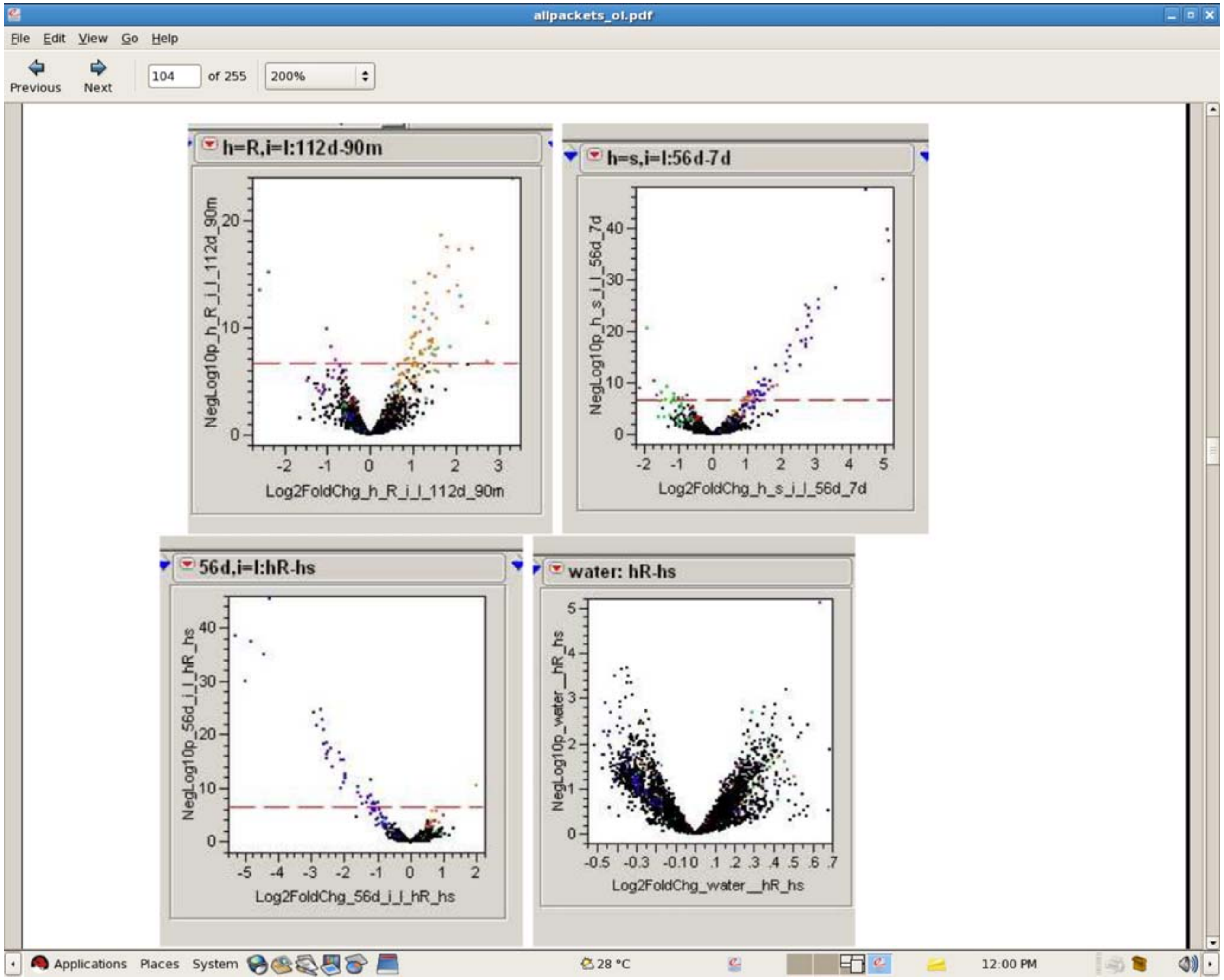
$$FDR = E\left(\frac{\text{Falsely rejected null hypotheses)}}{\text{Number of rejected null hypotheses}}\right)$$

See `qvalue()` in R and

`http://www4.stat.ncsu.edu/~jaosborn/research/microarray/software/qvalues.sas`

<u>A context in which multiplicity is a big issue:</u>

Microarray experiments, which may involve thousands of genes and tests



(Data courtesy of Cassi Myburg)

Bonferroni

Consider k contrasts of interest. Bonferroni adjustment to α which controls *fwe* is

$$\alpha' = \frac{\alpha}{k}$$

Simultaneous 95% confidence intervals for the k contrasts given by

$$a_1\bar{Y}_{1+} + a_2\bar{Y}_{2+} + \cdots + a_t\bar{Y}_{t+} \pm t(\frac{\alpha'}{2}, \nu)\sqrt{MS[E]\sum \frac{a_j^2}{n_j}}$$

and

$$b_1\bar{Y}_{1+} + b_2\bar{Y}_{2+} + \cdots + b_t\bar{Y}_{t+} \pm t(\frac{\alpha'}{2}, \nu)\sqrt{MS[E]\sum \frac{b_j^2}{n_j}}$$

$$\vdots$$

$$k_1\bar{Y}_{1+} + k_2\bar{Y}_{2+} + \cdots + k_t\bar{Y}_{t+} \pm t(\frac{\alpha'}{2}, \nu)\sqrt{MS[E]\sum \frac{k_j^2}{n_j}}$$

where ν denotes *df* for error. $t(\frac{\alpha'}{2}, \nu)$ might have to be obtained using software.

For the binding fraction example, consider only pairwise comparisons with Penicillin:

$$\theta_1 = \mu_1 - \mu_2, \theta_2 = \mu_1 - \mu_3, \theta_3 = \mu_1 - \mu_4, \theta_4 = \mu_1 - \mu_5$$

We have $k = 4$, $\alpha' = 0.05/k = 0.0125$, and $t(\frac{\alpha'}{2}, 15) =$ ________.

Substitution leads to

$$t(\alpha', 15)\sqrt{MS[E]\left(\frac{(-1)^2}{4} + \frac{(-1)^2}{4} + \frac{0^2}{4} + \cdots \frac{0^2}{4}\right)} = 2.84\sqrt{(9.05)\frac{2}{4}} = 6.0$$

so that $\boxed{\text{simultaneous}}$ 95% confidence intervals for $\theta_1, \theta_2, \theta_3, \theta_4$ take the form

$$\bar{y}_1 - \bar{y}_i \pm 6.0$$

```
proc glm data=one;     *In SAS, adjustment for k=4 achieved with care;
   title "Bonferroni correction for 4 contrasts";
   class drug;
   model  y=drug/clparm alpha=.0125;
   estimate "theta1" drug -1 1;
   estimate "theta2" drug -1 0 1;
   estimate "theta3" drug -1 0 0 1;
   estimate "theta4" drug -1 0 0 0 1;
run;
```

```
Bonferroni correction for 4 contrasts
The GLM Procedure

                                     Standard
Parameter               Estimate        Error    t Value    Pr > |t|     98.75% Confidence Limits

theta1                 2.7750000   2.12777270       1.30      0.2118     -3.2606985     8.8106985
theta2               -20.7750000   2.12777270      -9.76      <.0001    -26.8106985   -14.7393015
theta3                -9.5250000   2.12777270      -4.48      0.0004    -15.5606985    -3.4893015
theta4                -0.8000000   2.12777270      -0.38      0.7122     -6.8356985     5.2356985
```

(actually simultaneous **95**% confidence intervals)

Tukey

Tukey's method better than Scheffé 's method for $\boxed{\text{all pairwise}}$ comparisons in balanced designs Is conservative, controlling experimentwise error rate, and has lower type II error rate in these cases than Scheffé . (More powerful.)

For simple contrasts of the form

$$\theta = \mu_j - \mu_k$$

to test

$$H_0 : \theta = 0 \text{ vs } H_1 : \theta \neq 0$$

reject H_0 at level α if

$$|\hat{\theta}| > q(t, N-t, \alpha)\sqrt{\frac{MS[E]}{n}}$$

where $q(t, N-t, \alpha)$ denotes α level *studentized range* for t means and $N-t$ degrees of freedom. These studentized ranges can be found in Table C.11 of Rao.

For the IV data, $q(3, 15, 0.05) = 3.67$. Tukey's 95% honestly significant difference (HSD) for pairwise comparisons of treatment means in this balanced design are

$$3.67\sqrt{\frac{MS(E)}{n}} = \underline{\hspace{4cm}} = \underline{\hspace{4cm}}$$

```
proc glm;
   class firm;
   model con=firm;
   means firm/scheffe tukey;
run;
```

```
               Tukey's Studentized Range (HSD) Test for con

  NOTE: This test controls the Type I experimentwise error rate, but it
           generally has a higher Type II error rate than REGWQ.

               Alpha                                       0.05
               Error Degrees of Freedom                      15
               Error Mean Square                       9783.578
               Critical Value of Studentized Range      3.67338
               Minimum Significant Difference            148.33

        Means with the same letter are not significantly different.

           Tukey Grouping          Mean      N     firm

                        A        396.67      6     McGaw
                   B    A        273.83      6     Cutter
                   B             204.50      6     Abbott
```

(Scheffé excluded)

Expected mean squares

Definition: The $\boxed{\text{treatment mean square}}$ is given by

$$MS[Trt] = \frac{SS[Trt]}{t-1} = \frac{1}{t-1}\sum_i \sum_j (\bar{y}_{i+} - \bar{y}_{..})^2$$

$(\bar{y}_{..} = \overline{y}_{++}$ and $\bar{y}_{i+} = \frac{1}{n_i}\sum_{i=1}^{n_i} y_{ij})$

$$\boxed{\begin{aligned} E[MS[Trt]; H_1] &= E[SS[Trt]/(t-1); H_1] \\ &= \sigma^2 + \frac{1}{t-1}\sum n_i(\mu_i - \mu)^2 \\ &= \sigma^2 + n\frac{1}{t-1}\sum(\mu_i - \mu)^2 (\text{ balanced case}) \\ &= \sigma^2 + n\psi_T^2 \end{aligned}}$$

$$\text{where} \quad \psi_T^2 = \frac{1}{t-1}\sum(\mu_i - \mu)^2.$$

Note that under $H_0 : \mu_i \equiv \mu$ and $\psi_T^2 = 0$ so that

$$E[MS[Trt]; H_0] = E[SS[Trt]/(t-1); H_0] = \underline{\hspace{3cm}}$$

$\boxed{MS[E] = \frac{SS[E]}{N-t}}$, (generalization of pooled variance S_p^2 to $t > 2$ groups):

$$\begin{aligned}
MS[E] &= \frac{SS[E]}{N-t} = \frac{1}{N-t}\sum_{i=1}^{i=t}\sum_{j=1}^{j=n_i}(y_{ij} - \bar{y}_{i+})^2 \\
&= \frac{1}{N-t}\sum_{i=1}^{t}(n_i - 1)s_i^2 \\
&= \left(\frac{n_1 - 1}{N-t}\right)s_1^2 + \left(\frac{n_2 - 1}{N-t}\right)s_2^2 + \cdots + \left(\frac{n_t - 1}{N-t}\right)s_t^2 \\
&= \text{``}S_p^2\text{''}
\end{aligned}$$

Since $E(S_i^2) = \sigma^2$, $MS[E]$ is unbiased for σ^2 regardless of H_0 or H_1:

$$\begin{aligned}
E(S_i^2) &= \sigma^2 \Longrightarrow \\
E[MS[E]] &= \left(\frac{n_1 - 1}{N-t}\right)\sigma^2 + \left(\frac{n_2 - 1}{N-t}\right)\sigma^2 + \cdots + \left(\frac{n_t - 1}{N-t}\right)\sigma^2 \\
&= \sigma^2
\end{aligned}$$

$$\boxed{E[MS[E]] = \sigma^2}$$

Sample size computations for one-way ANOVA

Consider designing a completely randomized experiment that will have significance level α, power $1-\beta$, and sample size n to accept or reject the following hypotheses regarding the means of a response variable, Y with error variance σ^2:

$$H_0 : \mu_1 = \cdots = \mu_t \quad \text{vs} \quad H_a : \psi_T^2 = \frac{1}{t-1}\sum(\mu_i - \mu)^2.$$

Linear model, i.i.d. normal errors $\rightarrow$ can calculate any one quantity given others.
With H_0 true, $F = MS(Trt)/MS(E)$ follows an F-distribution under
With H_a true, $F = MS(Trt)/MS(E)$ follows a non-central F-distribution with non-centrality parameter given below ($\tau_i = \mu_i - \overline{\mu}$):

$$\gamma =$$

- Suppose $t = 4, n = 9(N = 36), \sigma^2 = 9, \alpha = .05$ and the hypotheses are

$$H_0 : \mu_1 = \cdots = \mu_4 \quad \text{vs} \quad H_a : \mu_1 = \mu_2 = 9, \mu_3 = 10, \mu_4 = 12.$$

Calculate the power, $P(\text{reject } H_0 | H_a \text{true})$. (an area under non-central F density).

```
> my.ncp <- 9*3*var(c(9,9,10,12))/9
> 1-pf(qf(.95,3,32),3,32,my.ncp)
[1] 0.4655894
```

Another example: consider these hypothese for antibiotic binding fractions:

$$H_1 : \mu_P = \mu + 3, \mu_T = \mu + 3, \mu_S = \mu - 6, \mu_E = \mu, \mu_C = \mu$$

Assume $\sigma = 3$ and we need to use $\alpha = \beta = 0.05$.

$$\gamma = n[(\frac{3}{3})^2 + (\frac{3}{3})^2 + (\frac{-6}{3})^2].$$

The following code should do the trick to calculate the necessary n

```
data one;
   do n=2 to 10;
      t=5; nu1=t-1; nu2=t*(n-1);
      sumtau2=3**2+3**2+(-6)**2;
      sigma2=9;
      ncp=n*sumtau2/sigma2;
      qf=finv(0.95,nu1,nu2);
      pf=probf(qf,nu1,nu2,ncp);
      power=1-pf;
      output;
   end;
run;
proc print;run;
```

OBS	N	T	NU1	NU2	SUMTAU2	SIGMA2	NCP	QF	PF	POWER
1	2	5	4	5	54	9	12	5.19217	0.59246	0.40754
2	3	5	4	10	54	9	18	3.47805	0.22465	0.77535
3	4	5	4	15	54	9	24	3.05557	0.06437	0.93563
4	5	5	4	20	54	9	30	2.86608	0.01533	0.98467
5	6	5	4	25	54	9	36	2.75871	0.00319	0.99681
6	7	5	4	30	54	9	42	2.68963	0.00060	0.99940

(not needed)