Linear regression using ordinary least squares

Malachy Campbell

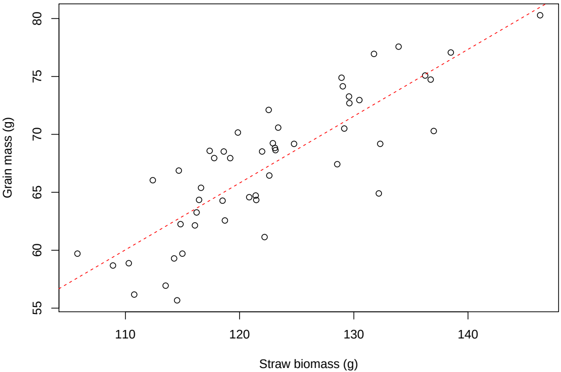
Sept. 4, 2020

OLS recap

Goal is to find a line that best fits the data

- Best line is one that minimizes the residual sum of squares
- Coefficients for best line are given by $\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

$$y_i = \beta_0 + \beta_1 x_i + e_i$$



OLS recap

We are partitioning variability in y into orthogonal components.

$$SS_{Total} = SS_{Regression} + SS_{Error}$$

Total variability in y = Variability explained by the model + Unexplained variability

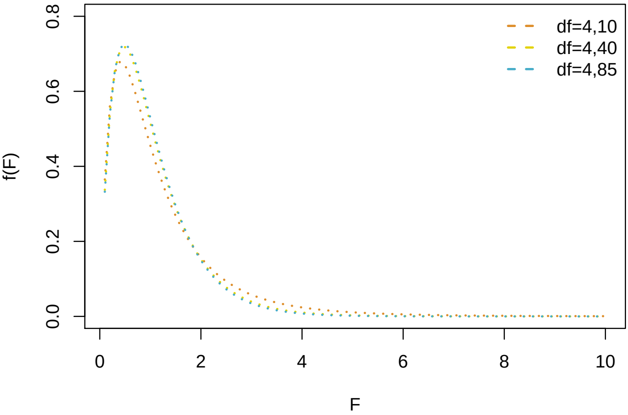
Source	**SS**	**df**	**MS**
Regression	$SSR = \sum_{i=1}^{i}(\hat{y}_i - \bar{y})^2$	$df_{Reg.} = p$	$\frac{SSR}{df_{Reg.}}$
Error	$SSE = \sum_{i=1}^{i}(y_i - \hat{y}_i)^2$	$df_{Err.} = n - p - 1$	$\frac{SS_{Err.}}{df_{Err.}}$
Total	$SS_{Tot.} = \sum_{i=1}^{i}\sum_{j=1}^{j}(y_i - \bar{y})^2$	$df_{Tot.} = n - 1$	

OLS recap

Does the model explain more variability than error?

- $\mathbf{H_0}$: $\beta_1 = 0$
- $\mathbf{H_1}$: $\beta_1 \neq 0$

$$F = \frac{\text{MSR}}{\text{MSE}}$$



F-dist. is parameterized by $\text{df}_{\textbf{Regression mod.}}$ and $\text{df}_{\textbf{Error}}$

Multiple linear regression – OLS with $p > 1$

Motivations:

- Confounding
  - Confounder: An extraneous variable that distorts the association between the dependent variable of interest and the response
- Gain precision
- Scientific reasons

An example

We are interested in looking at the effects of straw biomass (g plant^{-1}) and grain width (mm) on yield (g plant^{-1}) in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass, grain width and yield were collected.

An example

We are interested in looking at the effects of straw biomass (g plant^{-1}) and grain width (mm) on yield (g plant^{-1}) in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass, grain width and yield were collected.

- Assume grain width and plant biomass are uncorrelated.

	Biomass	**Gr. Width**	**Yld.**
Biomass	1.00	0.00	0.73
Gr. Width	0.00	1.00	0.67
Yld.	0.73	0.67	1.00

An example

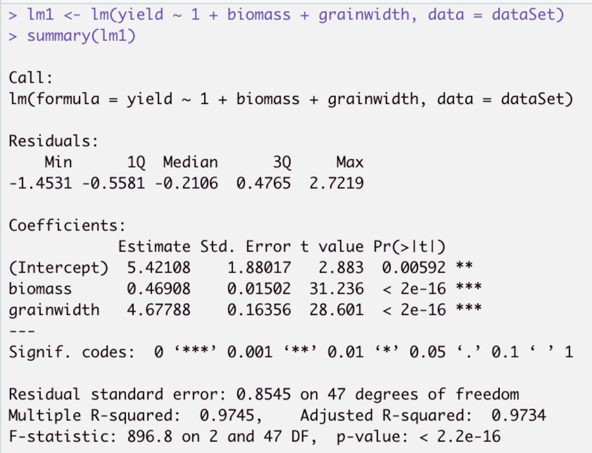
We are interested in looking at the effects of straw biomass (g plant^{-1}) and grain width (mm) on yield (g plant^{-1}) in rice. Fifty varieties were randomly selected and grown in the field. At harvest plant biomass, grain width and yield were collected.

- Assume grain width and plant biomass are uncorrelated.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

- y_i: grain yield for *i*th experimental unit (plant)
- x_{i1}: biomass for *i*th experimental unit (plant)
- x_{i2}: grain width for *i*th experimental unit (plant)

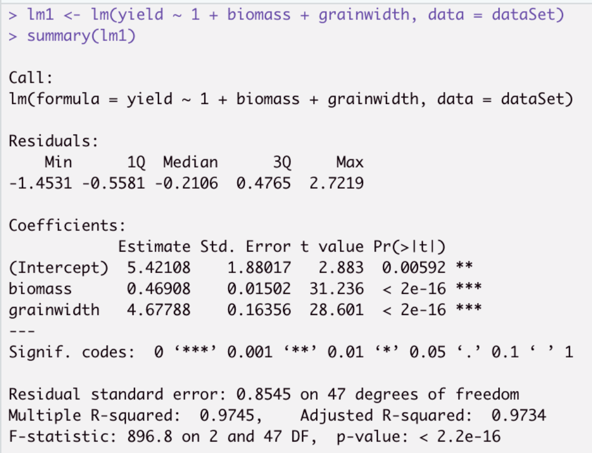
An example in R



```
> lm1 <- lm(yield ~ 1 + biomass + grainwidth, data = dataSet)
> summary(lm1)

Call:
lm(formula = yield ~ 1 + biomass + grainwidth, data = dataSet)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4531 -0.5581 -0.2106  0.4765  2.7219 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.42108    1.88017   2.883  0.00592 ** 
biomass      0.46908    0.01502  31.236  < 2e-16 ***
grainwidth   4.67788    0.16356  28.601  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8545 on 47 degrees of freedom
Multiple R-squared:  0.9745,	Adjusted R-squared:  0.9734 
F-statistic: 896.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

- Interpreting coefficients: If straw biomass is *held constant*, for every mm increase in grain width grain will result in a 4.7 g increase in yield.

An example in R



```
> lm1 <- lm(yield ~ 1 + biomass + grainwidth, data = dataSet)
> summary(lm1)

Call:
lm(formula = yield ~ 1 + biomass + grainwidth, data = dataSet)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4531 -0.5581 -0.2106  0.4765  2.7219 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.42108    1.88017   2.883  0.00592 ** 
biomass      0.46908    0.01502  31.236  < 2e-16 ***
grainwidth   4.67788    0.16356  28.601  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8545 on 47 degrees of freedom
Multiple R-squared:  0.9745,    Adjusted R-squared:  0.9734 
F-statistic: 896.8 on 2 and 47 DF,  p-value: < 2.2e-16
```

- The F-stat tells us the regression model explains sign. more variability in y than the error.

Model comparisons

Does a subset of predictors have an effect on the response?

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$$

- H_0: $\beta_3 = \beta_4 = 0$
- H_1: β_3 or $\beta_4 \neq 0$

Model comparisons

Does a subset of predictors have an effect on the response?

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$$

- H_0: $\beta_3 = \beta_4 = 0$
- H_1: β_3 or $\beta_4 \neq 0$

An alternative way to view this is as a set of nested models.

- H_0: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$ (reduced model)
- H_1: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$ (full model)

Model comparisons

- H_0: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$ (reduced model)
- H_1: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$ (full model)

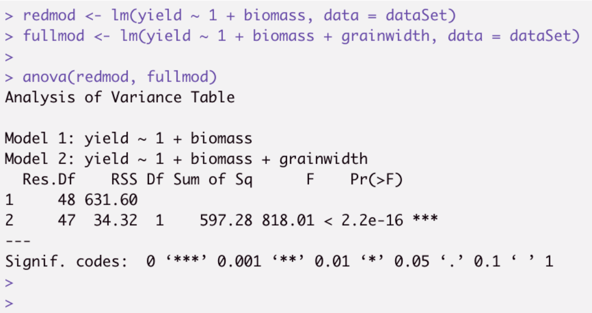
1. Fit “full” and “reduced” models and compute the residual/error SS ($\text{SS}_{\mathrm{Err.}}$).
2. Compare $\text{SS}_{\mathrm{Err.}}$ (SSE) between models using F-test

$$\mathrm{F} = \frac{(\mathrm{SSE}(\mathrm{Reduced}) - \mathrm{SSE}(\mathrm{Full}))/(df_{\mathrm{Reduced}} - df_{\mathrm{Full}})}{\mathrm{SSE}(\mathrm{Full})/df_{\mathrm{Full}}}$$

$$\mathrm{F}(df_{\mathrm{Reduced}} - df_{\mathrm{Full}}, df_{\mathrm{Full}})$$

Model comparisons in R

- The rice example only includes two predictors (biomass and grain width), but the concepts are the same.



```
> redmod <- lm(yield ~ 1 + biomass, data = dataSet)
> fullmod <- lm(yield ~ 1 + biomass + grainwidth, data = dataSet)
>
> anova(redmod, fullmod)
Analysis of Variance Table

Model 1: yield ~ 1 + biomass
Model 2: yield ~ 1 + biomass + grainwidth
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     48 631.60
2     47  34.32  1    597.28 818.01 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
```

Collinearity and multicollinearity

In many real world datasets there will be some dependency among predictors

- Issues:
  - Adding/changing predictor substantially changes coefficient estimates
  - Can no longer interpret meaning of coefficients
  - High standard error for coefficient estimates

- Addressing dependencies
  - Correlation between predictors
  - Variance inflation factors (VIF)

Variance inflation factors

Regress each predictor variable on all other predictors and calculate R^2 **for each predictor**

For example:

- Our regression model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$
- For $\mathbf{x_1}$: $x_{i1} = \alpha + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} + e_i$
- VIF for $\mathbf{x_1}$:

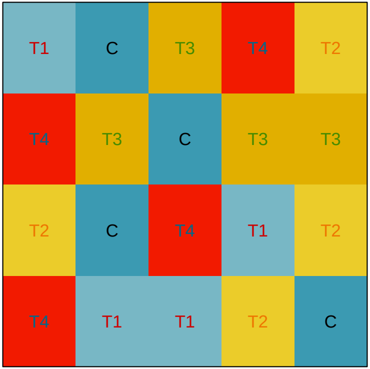
$$VIF_1 = \frac{1}{1 - R_1^2}$$

- Remove predictors with $VIF > 10$

Analyzing one factor designs

Completely randomized designs

Objective: One factor (e.g. treatment) under consideration and levels of the factor are randomly assigned to experimental units

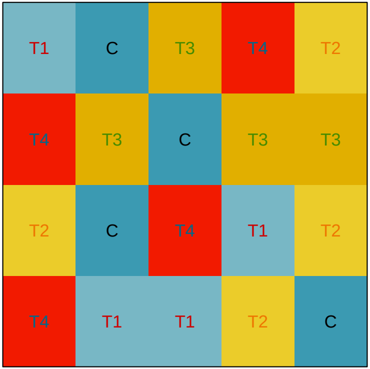


T1	C	T3	T4	T2
T4	T3	C	T3	T3
T2	C	T4	T1	T2
T4	T1	T1	T2	C

Randomization ensures that the systematic difference is treatment (i.e. eliminates confounding)

Completely randomized designs

Objective: One factor (e.g. treatment) under consideration and levels of the factor are randomly assigned to experimental units



T1	C	T3	T4	T2
T4	T3	C	T3	T3
T2	C	T4	T1	T2
T4	T1	T1	T2	C

The model: $y_{ij} = \mu_i + e_{ij}$

Comparing treatments (means)

Two levels: two-sample t-test

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{2/n}}$$

More than two levels?

Comparing treatments (means)

Two levels: two-sample t-test

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}\sqrt{2/n}}$$

More than two levels: One-way analysis of variance (ANOVA)

- The ANOVA model is often written as $y_{ij} = \mu_i + e_{ij}$
  - y_{ij}: the response for the jth replicate in the ith treatment
  - μ_i: mean of the ith treatment
  - **Cell means model**
- An equivalent model is $y_{ij} = \mu + \alpha_i + e_{ij}$
  - α_i: ith treatment effect (how far the ith treatment deviates from the overall mean)
  - **Additive model**

Example – Single factor CRD

Researchers are interested in studying the effects of three N regimes (180, 200, and 220 lbs. N ac^{-1}) on yield. The field was split into nine plots (experimental units) and treatments were randomly assigned to each plot. Three observations were recorded for each combination of N level ($N = 3 \times 3 = 9$).

Example – Single factor CRD

Researchers are interested in studying the effects of three N regimes (180, 200, and 220 lbs. N ac^{-1}) on yield. The field was split into nine plots (experimental units) and treatments were randomly assigned to each plot. Three observations were recorded for each combination of N level ($N = 3 \times 3 = 9$).

- **Treatment design:** Each Nitro. treatment is assigned to at least one experimental unit
  - One factor: Nitrogen
  - Three levels: 180, 200, and 220
- **Experimental Design:** Treatments are randomly assigned and all are observed an equal number of times → *balanced complete randomized design*

Example – Single factor CRD

Researchers are interested in studying the effects of three N regimes (180, 200, and 220 lbs. N ac^{-1}) on yield. The field was split into nine plots (experimental units) and treatments were randomly assigned to each plot. Three observations were recorded for each combination of N level ($N = 3 \times 3 = 9$).

$$y_{ij} = \mu_i + e_{ij}$$

$\mathbf{H_0}$: Average yield is the same for all Nit.; **$\mathbf{H_1}$**: Average yield is different for at least one comparison of N levels;

$\mathbf{H_0}$: $\mu_{20} = \mu_{25} = \mu_{30}$; **$\mathbf{H_1}$**: $\mu_i \neq \mu_k$

Example – Single factor CRD

Researchers are interested in studying the effects of three N regimes (180, 200, and 220 lbs. N ac^{-1}) on yield. The field was split into nine plots (experimental units) and treatments were randomly assigned to each plot. Three observations were recorded for each combination of N level ($N = 3 \times 3 = 9$).

$$y_{ij} = \mu + \alpha i + e_{ij}$$

$$\mathbf{H_0}\text{: } \alpha_i = 0\text{; } \mathbf{H_1}\text{: } \alpha_i \neq 0$$

ANOVA is just OLS

ANOVA is regression on dummy variables

- Dummy variables: recoding of categorical variables

Cell means: $y_{ij} = \mu_i + e_{ij}$

Nit.	Rep.	Yld.
180	1	173.3
180	2	182.9
180	3	169.6
200	1	205.9
200	2	208.5
200	3	203.9
220	1	229.1
220	2	231.3
220	3	208.7

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

ANOVA is just OLS

ANOVA is regression on dummy variables

- Dummy variables: recoding of categorical variables

Nit.	Rep.	Yld.
180	1	173.3
180	2	182.9
180	3	169.6
200	1	205.9
200	2	208.5
200	3	203.9
220	1	229.1
220	2	231.3
220	3	208.7

Additive: $y_{ij} = \mu + \alpha_i + e_{ij}$

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

An example in R

Researchers are interested in studying the effects of three N regimes (180, 200, and 220 lbs. N ac^{-1}) on yield. The field was split into nine plots (experimental units) and treatments were randomly assigned to each plot. Three observations were recorded for each combination of N level ($N = 3 \times 3 = 9$).

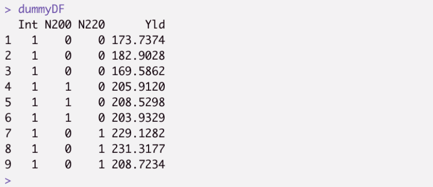
$$\mathbf{H_0}\text{: } \alpha_i = 0\text{; } \mathbf{H_1}\text{: } \alpha_i \neq 0$$

OR

$$\mathbf{H_0}\text{: } \mu_{20} = \mu_{25} = \mu_{30}\text{; } \mathbf{H_1}\text{: } \mu_i \neq \mu_k$$

An example in R

Using dummy variables:

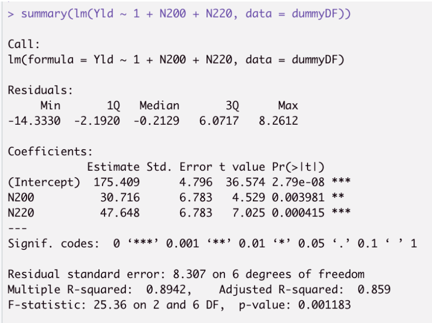


```
> dummyDF
  Int N200 N220      Yld
1   1    0    0 173.7374
2   1    0    0 182.9028
3   1    0    0 169.5862
4   1    1    0 205.9120
5   1    1    0 208.5298
6   1    1    0 203.9329
7   1    0    1 229.1282
8   1    0    1 231.3177
9   1    0    1 208.7234
>
```

Using `aov()`:

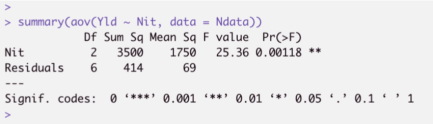
An example in R

Using dummy variables:



```
> summary(lm(Yld ~ 1 + N200 + N220, data = dummyDF))

Call:
lm(formula = Yld ~ 1 + N200 + N220, data = dummyDF)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.3330  -2.1920  -0.2129   6.0717   8.2612 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  175.409      4.796  36.574 2.79e-08 ***
N200          30.716      6.783   4.529 0.003981 ** 
N220          47.648      6.783   7.025 0.000415 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.307 on 6 degrees of freedom
Multiple R-squared:  0.8942,	Adjusted R-squared:  0.859 
F-statistic: 25.36 on 2 and 6 DF,  p-value: 0.001183
```

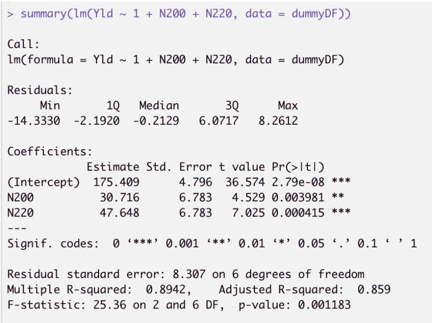
Using `aov()`:



```
> 
> summary(aov(Yld ~ Nit, data = Ndata))
            Df Sum Sq Mean Sq F value  Pr(>F)   
Nit          2   3500    1750   25.36 0.00118 **
Residuals    6    414      69                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

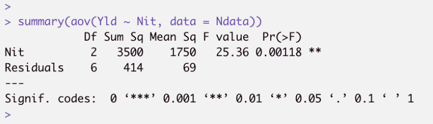
An example in R

Using dummy variables:



```
> summary(lm(Yld ~ 1 + N200 + N220, data = dummyDF))

Call:
lm(formula = Yld ~ 1 + N200 + N220, data = dummyDF)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.3330  -2.1920  -0.2129   6.0717   8.2612 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  175.409      4.796  36.574 2.79e-08 ***
N200          30.716      6.783   4.529 0.003981 ** 
N220          47.648      6.783   7.025 0.000415 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.307 on 6 degrees of freedom
Multiple R-squared:  0.8942,	Adjusted R-squared:  0.859 
F-statistic: 25.36 on 2 and 6 DF,  p-value: 0.001183
```

Using `aov()`:

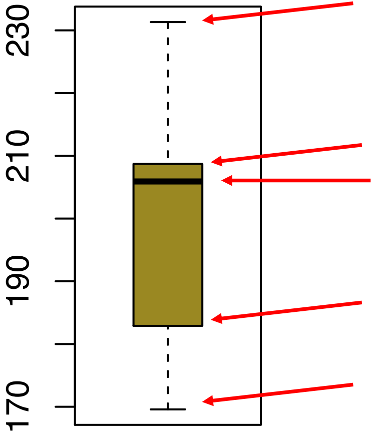


```
>
> summary(aov(Yld ~ Nit, data = Ndata))
            Df Sum Sq Mean Sq F value  Pr(>F)   
Nit          2   3500    1750   25.36 0.00118 **
Residuals    6    414      69                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

With ANOVA we are interested in drawing conclusions about a factor(s) (set of predictors)

- Equivalent to comparing models

Post-hoc tests



Q3 + 1.5*IQR

3rd quartile (Q3):
median of upper half

Median

1st quartile (Q1):
median of lower half

Q1 – 1.5*IQR

$$IQR = Q3 - Q1$$

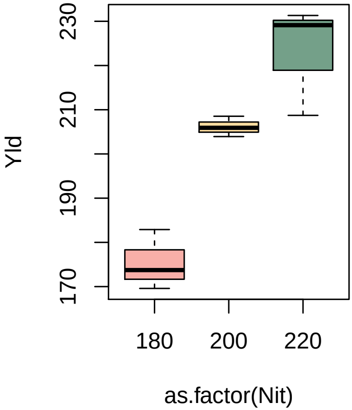
Post-hoc tests

Which treatments are different?

$\mathbf{H_0}$: $\alpha_i = 0$; $\mathbf{H_1}$: $\alpha_i \neq 0$

OR

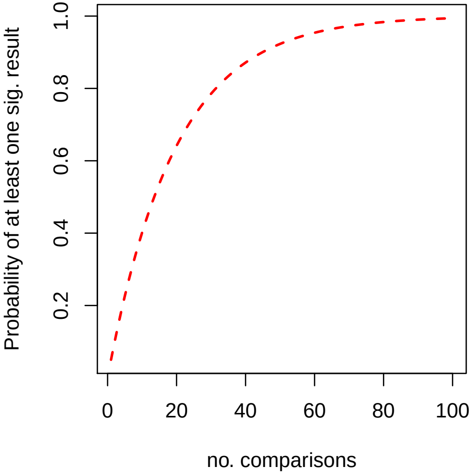
$\mathbf{H_0}$: $\mu_{20} = \mu_{25} = \mu_{30}$; $\mathbf{H_1}$: $\mu_i \neq \mu_k$



Hypothesis testing – multiple testing

We want to compare four levels of nitrogen. For each comparison we assume $\alpha = 0.05$

- $\frac{k(k-1)}{2}$ comparisons
- $1 - \alpha$: probability of failing to reject null when null is true
- Family-wise error rate: Type I error over set of comparisons
    - $FWER = 1 - (1-\alpha)^{\text{no. tests}}$

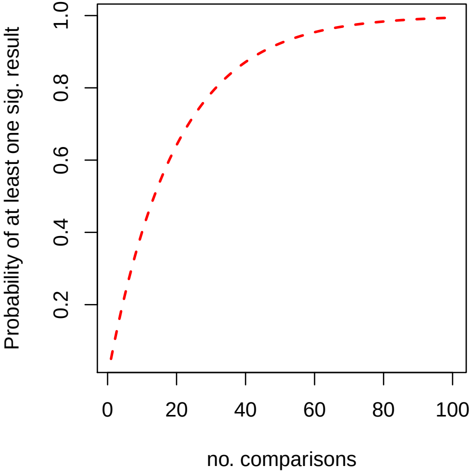


Hypothesis testing – multiple testing

We want to compare four levels of nitrogen. For each comparison we assume $\alpha = 0.05$

Multiple testing correction:

- Bonferroni's correction: $\alpha^* = \alpha/\text{No. tests}$
- Sidak's correction: $\alpha^* = 1-(1-\alpha)^{1/\text{No. tests}}$
- Benjamini-Hochberg: Ranking procedure (see course notes)

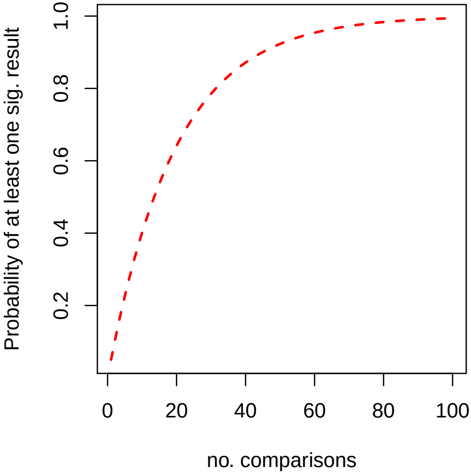


Hypothesis testing – multiple testing

We want to compare four levels of nitrogen. For each comparison we assume $\alpha = 0.05$

Post-hoc tests:

- Tukey's honest significant difference (HSD) test: all pairwise comp.
- Dunnett's test: Control vs treatments



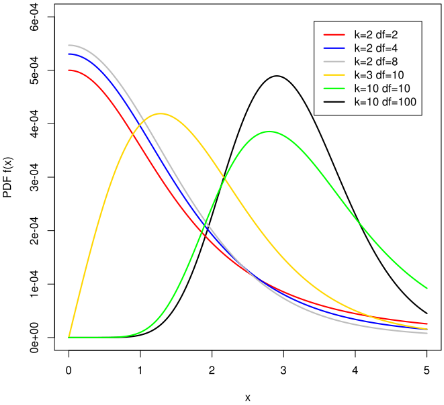
Hypothesis testing – Tukey's HSD

HSD stat. is similar to t-statistic:

$$HSD = \frac{\mu_1 - \mu_2}{\sqrt{MSE_e/n}}$$

$$\mu_1 \geq \mu_2$$

Studentized range distribution:



https://en.wikipedia.org/wiki/Studentized$_{r}$*ange*$_{d}$*istribution*

k = number of levels for factor, $df = n - k$

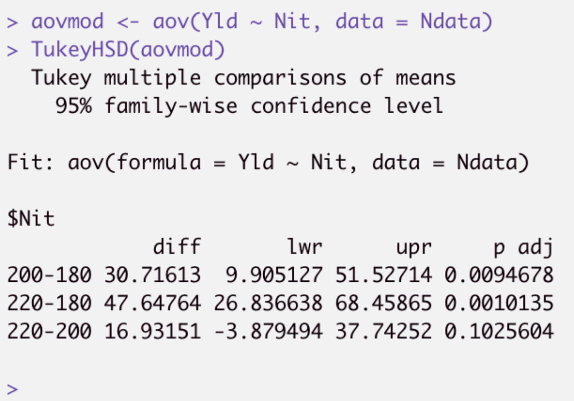
Hypothesis testing – Tukey's HSD

HSD stat. is similar to t-statistic:

$$HSD = \frac{\mu_1 - \mu_2}{\sqrt{MSE_e/n}}$$

$$\mu_1 \geq \mu_2$$

In R:



```
> aovmod <- aov(Yld ~ Nit, data = Ndata)
> TukeyHSD(aovmod)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Yld ~ Nit, data = Ndata)

$Nit
            diff       lwr      upr     p adj
200-180 30.71613  9.905127 51.52714 0.0094678
220-180 47.64764 26.836638 68.45865 0.0010135
220-200 16.93151 -3.879494 37.74252 0.1025604


>
```