ENV 710: Lecture 11

linear models 2

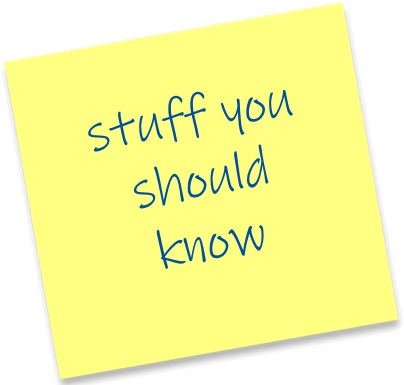
Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

linear models

categorical explanatory variables

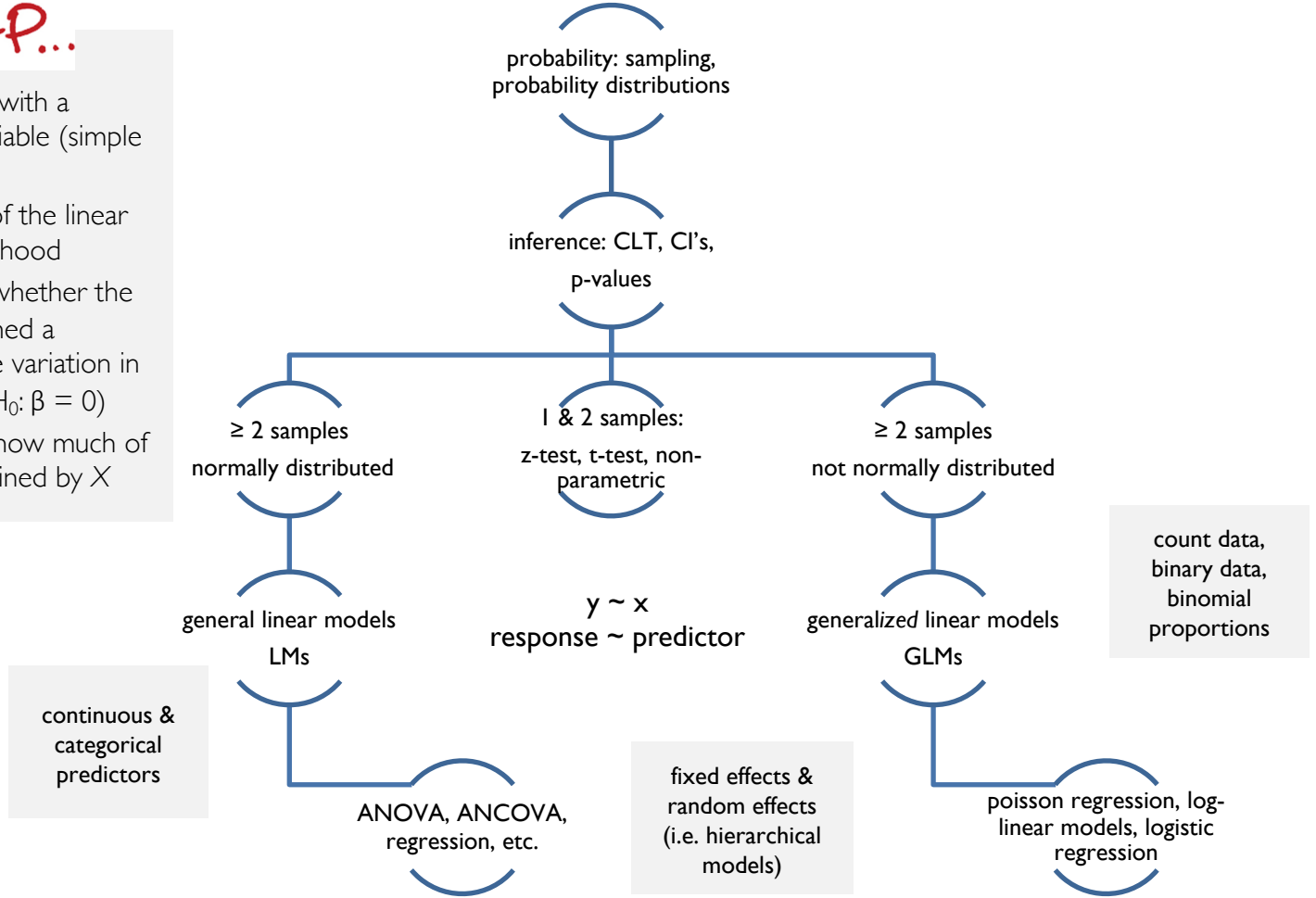
learning goals

- linear models with nominal explanatory variables
  - ANOVA: continuous response, nominal explanatory variables
  - assumptions
  - contrasts and treatment contrasts
- model validation techniques
- *post-hoc* multiple comparisons tests
- linear model with 2 or more nominal predictors



LET'S RECAP...

- we explored linear models with a continuous explanatory variable (simple linear regression)
- estimated the parameters of the linear model using maximum likelihood
- used ANOVA to estimate whether the explanatory variable explained a significant proportion of the variation in the response variable (i.e., H_0: $\beta = 0$)
- calculated the R^2 to assess how much of the variation in *Y* was explained by *X*

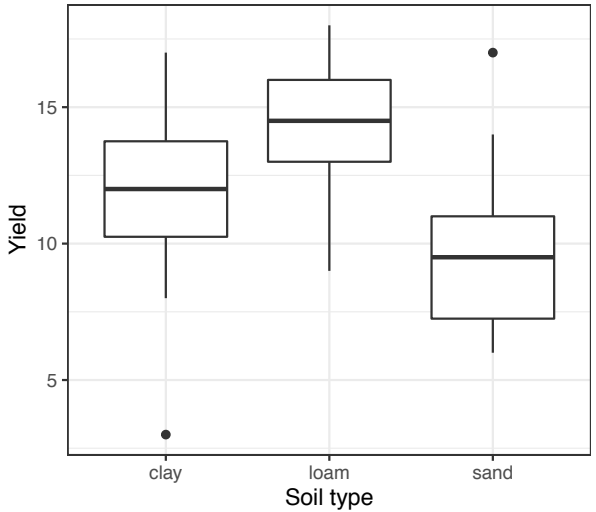


example

soil type

- crop yields per unit area measured from 10 randomly selected fields on 3 soil types: sand, clay, loam
- soil is a nominal variable with 3 levels
- null hypothesis: no soil type has an effect on yield

S	L	C	S	L	S	C	L	L	S	S	C	L	L	S
C	S	L	S	C	L	C	S	L	C	C	L	C	S	C



lm with nominal variable

- hypotheses:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

$$H_a : \text{Not } \mu_1 = \mu_2 = ... = \mu_k$$

- like the ANOVA conducted for regression, we partition the variability into different components

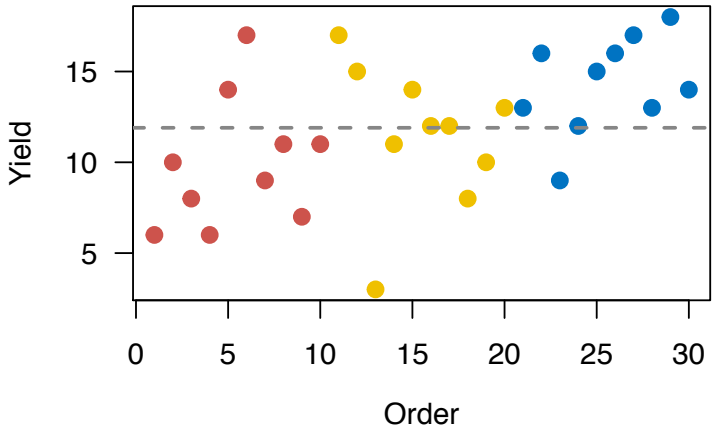
- partitioning variability – what portion of the variability is due to the variable of interest and what portion is due to other factors

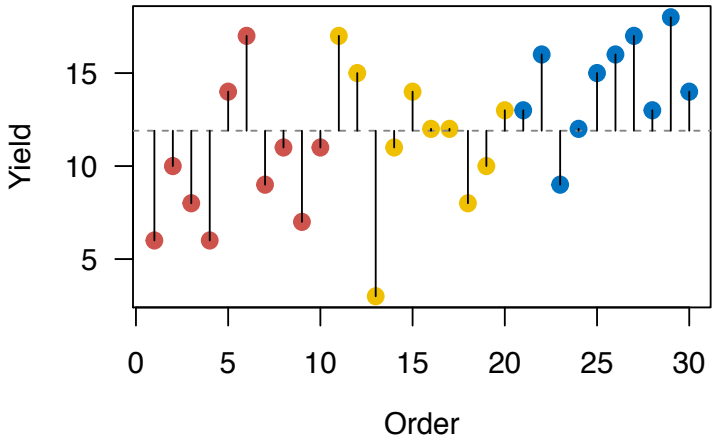
$$SST = SSA + SSE$$

$$SS_{\text{total}} = SS_{\text{among groups}} + SS_{\text{within groups}}$$

$$\sigma_Y^2 = \sigma_A^2 + \sigma^2$$

total sum of squares





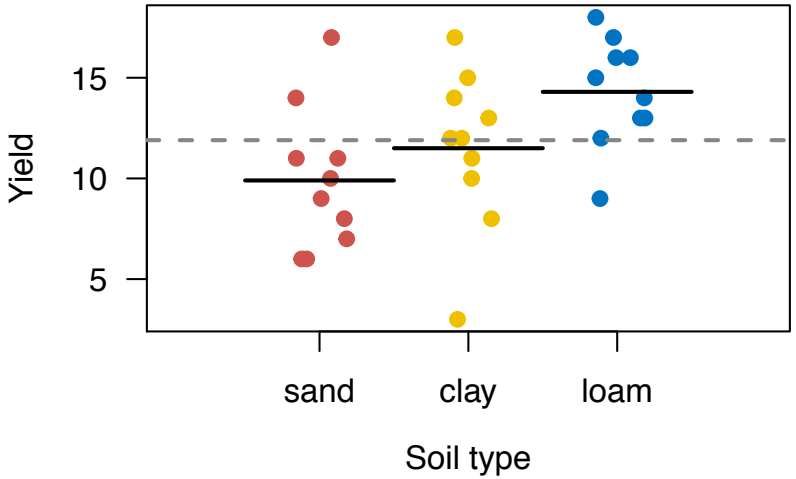
total sum of squares

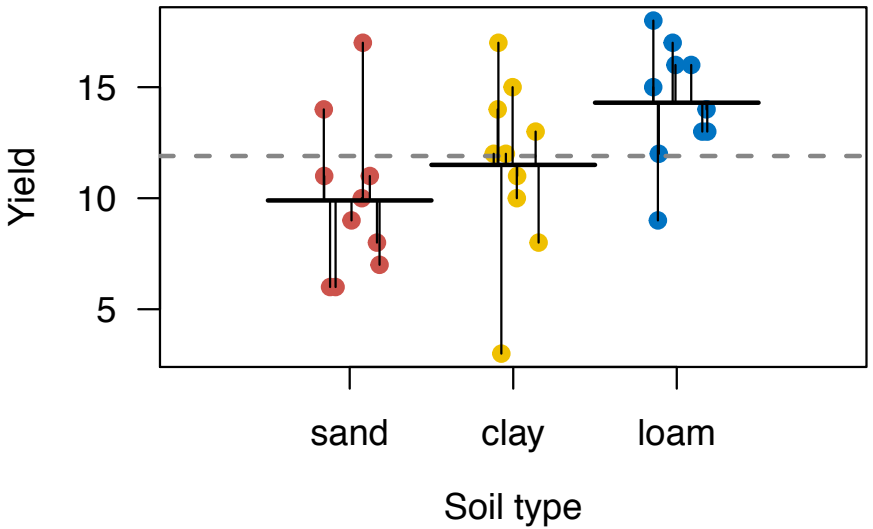
- SST is the sum of the squares of the lengths of the lines joining each data point to the overall mean
- SST measures the total variability in the response variable

$$SS_{\text{total}} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y})^2$$

k = no. of groups or levels
i = group counter
j = individual counter
n = observations per group
N = sample size
y = observation

error sum of squares



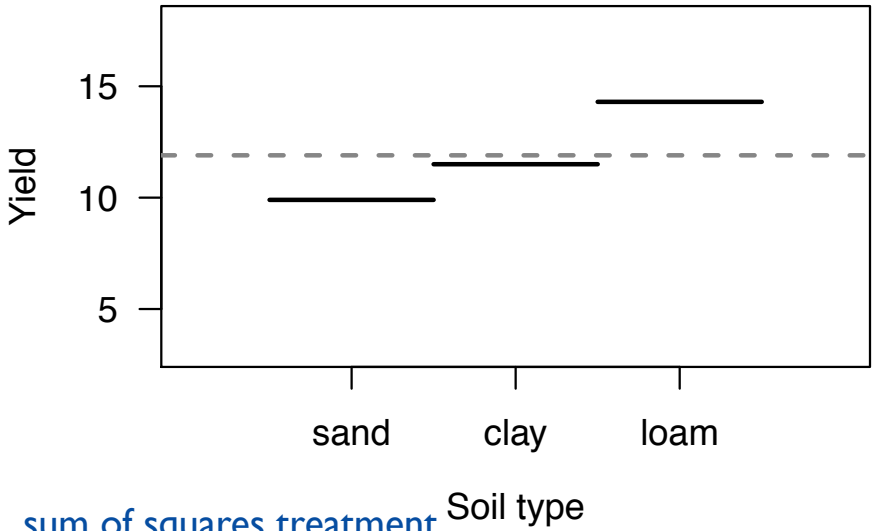


error sum of squares

- measures the variability within groups, variability unexplained by the group variable
- SSE is the sum of the squares of the length of the lines joining each data point to its treatment mean
- SSE is also called SS_{within}

$$SS_{\text{within}} = \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2$$

treatment sum of squares



$$SS_{\text{among}} = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$

sum of squares treatment

- measures the variability among groups
- explained variability: deviation of group mean from overall mean, weighted by sample size

because $SST = SSA + SSE$

- if SST is big and SSE is small, SSA is big
- if SST is big and SSE is big, SSA is small

linear models in R

- use function `lm()` or `aov()`
- same syntax as for linear regression

```
yield.lm <- lm(y ~ soil, data=yield)
```

- R detects that soil is a nominal explanatory variable with 3 levels

```
class(yield$soil)
```

- produce ANOVA table with sum of squares using `summary.aov()`

equivalent R statements

```
summary.aov(lm(y ~ soil, data=yield))
summary(aov(yield ~ factor(soil)))
anova(lm(yield ~ factor(soil)))
```

linear models in R

```
summary.aov(yield.lm)
             Df Sum Sq Mean Sq F value Pr(>F)
factor(soil)  2   99.2   49.60   4.245  0.025 *
Residuals    27  315.5   11.69
---
```

$$\mathrm{df}_{treat} = k - 1$$

$$\mathrm{df}_{resid} = k(n_i - 1)$$

$$\mathrm{df}_{tot} = n - 1$$

linear models in R

$$SS_{\text{among}} = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$

```
summary.aov(yield.lm)
             Df Sum Sq Mean Sq F value Pr(>F)
factor(soil)  2   99.2   49.60   4.245  0.025 *
Residuals    27  315.5   11.69
---
```

$$\text{df}_{treat} = k - 1$$
$$\text{df}_{resid} = k(n_i - 1)$$
$$\text{df}_{tot} = n - 1$$

$$SS_{\text{within}} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$$

sum of squares: measures of variability among and within groups

linear models in R

$$SS_{\text{among}} = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$

```
summary.aov(yield.lm)
             Df Sum Sq Mean Sq F value Pr(>F)
factor(soil)  2   99.2   49.60   4.245  0.025 *
Residuals    27  315.5   11.69
---
```

$$\text{MS} = \frac{\text{SS}}{\text{df}}$$

$$\text{df}_{treat} = k - 1$$
$$\text{df}_{resid} = k(n_i - 1)$$
$$\text{df}_{tot} = n - 1$$

$$SS_{\text{within}} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$$

mean squares: average variability among and within groups, calculated as the total variability (sum of squares) scaled by the associated degrees of freedom

linear models in R

F value is the ratio of among group and within group variability

$$SS_{\text{among}} = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$$

$$F = \frac{MSA}{MSE}$$

$$F = \frac{n\sigma_A^2 + \sigma^2}{\sigma^2}$$

```
summary.aov(yield.lm)
             Df Sum Sq Mean Sq F value Pr(>F)
factor(soil)  2   99.2   49.60   4.245  0.025 *
Residuals    27  315.5   11.69
---
```

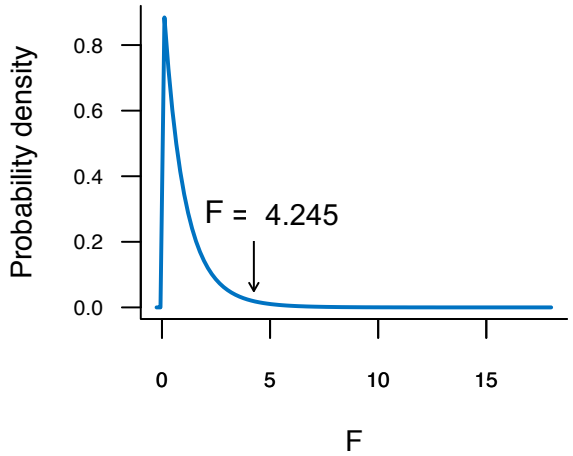
$$\text{MS} = \frac{\text{SS}}{\text{df}}$$

$$\text{df}_{treat} = k - 1$$
$$\text{df}_{resid} = k(n_i - 1)$$
$$\text{df}_{tot} = n - 1$$

$$SS_{\text{within}} = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2$$

linear models in R

```
summary.aov(yield.lm)
             Df Sum Sq Mean Sq F value Pr(>F)
factor(soil)  2   99.2   49.60   4.245  0.025 *
Residuals    27  315.5   11.69
---
```

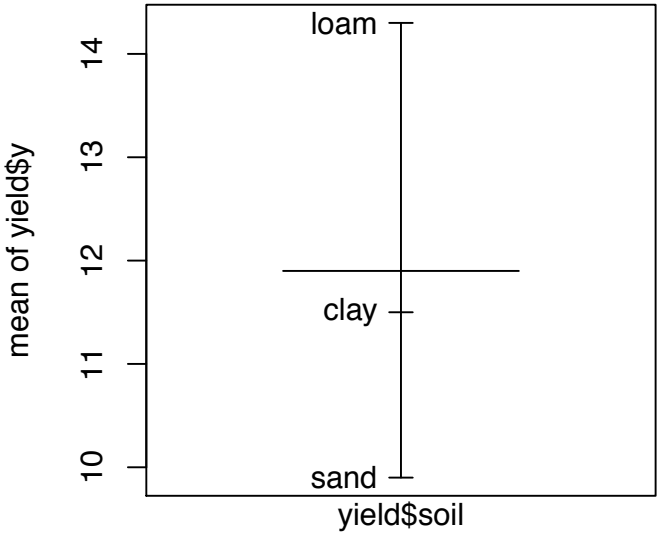


$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

$$H_a : \text{Not } \mu_1 = \mu_2 = ... = \mu_k$$

- if reject H_0: at least one pair of population means are different from each other
- if fail to reject H_0: no pairs of population means are different from one another; observed differences in sample means are attributable to sampling variability

linear models in R



- what about estimates of parameters – effect sizes?

```
plot.design(y~soil)
```

- loam has highest yield, followed by clay and sand

linear models in R

```
summary(yield.lm)

Call:
lm(formula = y ~ factor(soil), data = yield)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.500      1.081  10.638  3.7e-11 ***
factor(soil)loam    2.800      1.529   1.832   0.0781 .
factor(soil)sand   -1.600      1.529  -1.047   0.3046
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.418 on 27 degrees of freedom
Multiple R-squared:  0.2392,  Adjusted R-squared:  0.1829
F-statistic: 4.245 on 2 and 27 DF,  p-value: 0.02495
```

linear models in R

- effect sizes are shown in the form of contrasts so that the equation of a linear model can be built when the explanatory variable is nominal
- linear regression model is represented by

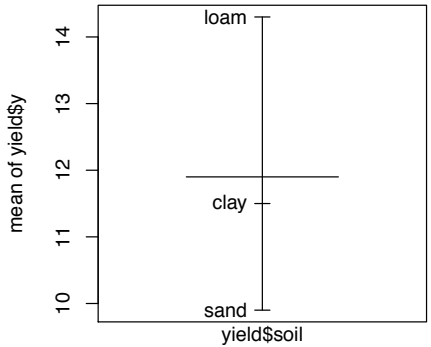
$$y = a + bx_1$$

- lm with a nominal explanatory variable with 3 levels

$$y = a + bx_{clay} + cx_{loam} + dx_{sand}$$

overall mean

b, *c*, *d* are differences between the overall mean and the means for a particular level



linear models in R

$$y = a + bx_{clay} + cx_{loam} + dx_{sand}$$

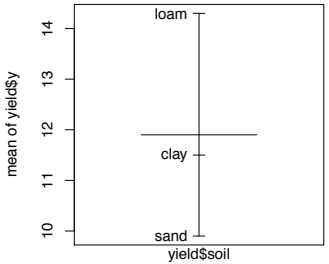
$$y_{sand} = a + (b \cdot 0) + (c \cdot 0) + d \cdot 1$$

$$\hat{y}_{sand} = a + d$$

here we are estimating 4 parameters when we only need 3 – there's a redundant parameter

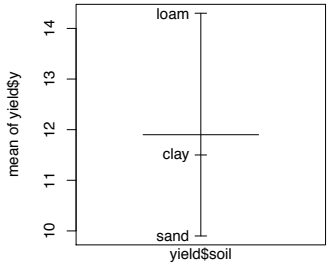
- R's default is to remove the overall mean parameter, *a*, and replace it with one of the treatment means (usually the first alphabetically) → treatment contrasts
- remaining parameters are the differences (contrasts) between this mean and the 2 other level means

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.500      1.081  10.638  3.7e-11 ***
factor(soil)loam    2.800      1.529   1.832   0.0781 .
factor(soil)sand   -1.600      1.529  -1.047   0.3046
---
```



linear models in R

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         11.500      1.081  10.638  3.7e-11 ***
factor(soil)loam     2.800      1.529   1.832   0.0781 .
factor(soil)sand    -1.600      1.529  -1.047   0.3046
---
```



- estimate of the intercept is the mean crop yield for soil type clay
- `(soil)loam` estimate is the difference between clay and loam

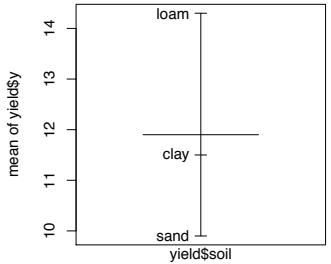
  11.5 + 2.8 = 14.3

- `(soil)`sand estimate is the difference between clay and sand

  11.5 – 1.6 = 9.9

linear models in R

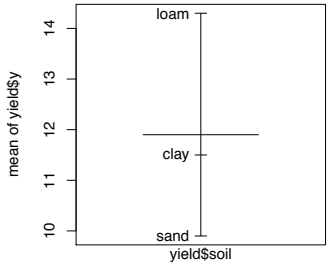
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         11.500      1.081  10.638  3.7e-11 ***
factor(soil)loam     2.800      1.529   1.832   0.0781 .
factor(soil)sand    -1.600      1.529  -1.047   0.3046
---
```



- standard errors provide unreliability of estimates
- standard error of the intercept is the standard error of a mean (clay)
- standard error of `(soil)loam` and `(soil)sand` is the standard error of the difference between two means

linear models in R

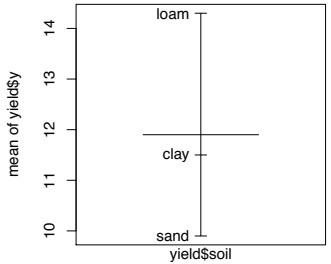
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         11.500      1.081  10.638  3.7e-11 ***
factor(soil)loam     2.800      1.529   1.832   0.0781 .
factor(soil)sand    -1.600      1.529  -1.047   0.3046
---
```



- t-value of intercept tests H_0: β=0
- t-value of `(soil)loam` tests H_0: β_{loam} - β_{clay} =0
- no significant difference in crop yield between loam and clay soil types
- similar interpretation for sand

linear models in R

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         11.500      1.081  10.638  3.7e-11 ***
factor(soil)loam     2.800      1.529   1.832   0.0781 .
factor(soil)sand    -1.600      1.529  -1.047   0.3046
---
```



- what about the difference between loam and sand?

$$H_0: \beta_{sand} - \beta_{loam} = 0$$

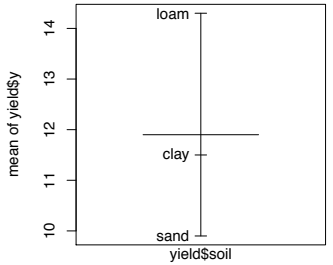
- several ways to do comparison…
- set baseline contrast to loam using `relevel()` and refit the model

```
yield$soil <- relevel(yield$soil, ref = "loam")
yield.lm2 <- lm(y ~ soil, data=yield)
```

linear models in R

```
yield.lm2 <- lm(y ~ soil, data=yield)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    14.300      1.081  13.229 2.58e-13 ***
soilclay       -2.800      1.529  -1.832  0.07807 .
soilsand       -4.400      1.529  -2.878  0.00773 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.418 on 27 degrees of freedom
Multiple R-squared:  0.2392, Adjusted R-squared:  0.1829
F-statistic: 4.245 on 2 and 27 DF,  p-value: 0.02495
```



- now `soilloam` is the baseline
- t-value for `soilsand` is large, so can reject H_0
- crop yield is significantly lower in sand than in loam

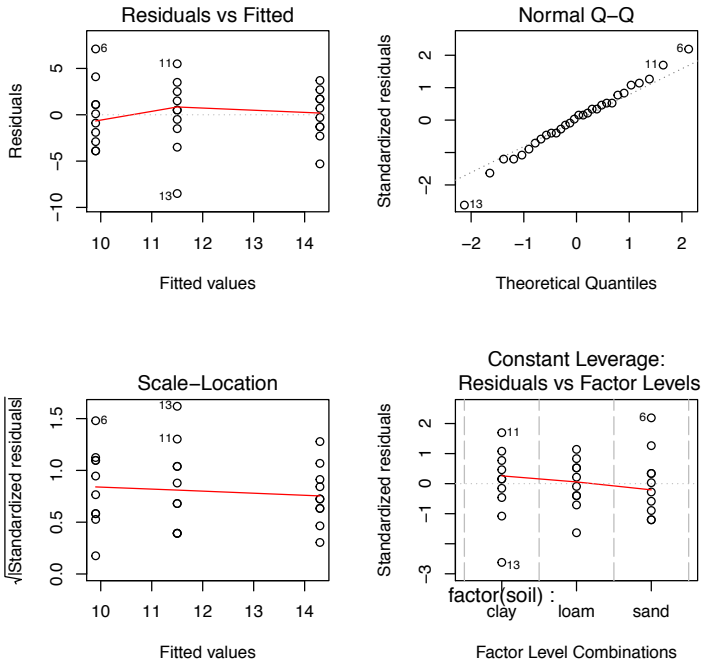
assumptions

- scores for each treatment level are normally distributed around their means
- independent observations (within and between groups)
- homogeneity of variances (homoscedasticity)
  - equal standard deviations of the populations
  - largest SD is less than twice the smallest SD

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$$

lm models in R

```
plot(yield.lm)
```



notes

- check assumptions before starting – can transform data if there are gross deviations from normality
- if no gross deviations, proceed and check assumptions after fitting the model using `plot()`
- residual plots assess homogeneity of variances, QQ plot assesses normality, leverage plot assesses influential data points

multiple comparisons

- there are other ways than refitting the model to determine if treatment level means are significantly different
- use multiple comparisons only after rejecting the overall H_0
- more likely to incorrectly reject the H_0 when dealing with many inferences
  - more likely to find low p-value by chance
  - may find differences between groups that really don't exist
- correction approaches:
  - Bonferroni
  - Tukey (equal group sizes)
  - Tukey-Kramer (unequal group sizes)

```
Call:
lm(formula = y ~ factor(soil), data = yield)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        11.500      1.081  10.638  3.7e-11
***
factor(soil)loam    2.800      1.529   1.832   0.0781 .
factor(soil)sand   -1.600      1.529  -1.047   0.3046
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

Residual standard error: 3.418 on 27 degrees of freedom
Multiple R-squared:  0.2392, Adjusted R-squared:  0.1829
F-statistic: 4.245 on 2 and 27 DF,  p-value: 0.02495
```

Tukey's HSD

- Tukeys' Honestly Significant Difference
- controls for the fact that multiple simultaneous comparisons are being made
- p-value adjusted downward to achieve an error rate of 0.05

$$HSD = q\sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) MS_{residual}}$$

- HSD is the difference between means necessary to have an α of 0.05

note: need to wrap `yield.lm` in `aov()` before running `TukeyHSD()`

```
TukeyHSD(aov(yield.lm))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield.lm)

$`factor(soil)`
            diff        lwr        upr     p adj
loam-clay    2.8 -0.9903777  6.5903777 0.1785489
sand-clay   -1.6 -5.3903777  2.1903777 0.5546301
sand-loam   -4.4 -8.1903777 -0.6096223 0.0204414
```

example mangrove

Are there positive effects of two common sponge species and location on root growth of the mangrove tree, *Rhizophora mangle?*

```
sponge <- read.csv("CompleteSpongeData.csv", sep=",", header=T)

class(sponge$Treatment)
[1] "factor"

levels(sponge$Treatment)
[1] "Control"   "Foam"      "Haliclona" "Tedania"

levels(sponge$Location)
[1] "bbs" "etb" "lcn" "lcs"
```

example mangrove

Are there positive effects of two common sponge species and location on root growth of the mangrove tree, *Rhizophora mangle?*

1. set up hypotheses and determine α level (usually 0.05)
2. determine the appropriate model
   - linear model with two nominal explanatory variables
3. preliminarily check whether data fit model assumptions
   - plot data, check homogeneity of variance and normality of data
4. run your model
5. validate the model assumptions
   - plot residuals, check homogeneity of variance and normality of data
6. make conclusions and interpret effects

example mangrove

Are there positive effects of two common sponge species and location on root growth of the mangrove tree, *Rhizophora mangle?*

```
sponge.lm <- lm(RootGrowthRate ~ Treatment + Location, data=sponge)
 summary.aov(sponge.lm)
```

```
                Df   Sum Sq    Mean Sq   F value    Pr(>F)
Treatment       3    4.402     1.4673    6.958      0.000393 ***
Location        3    0.806     0.2687    1.274      0.290658
Residuals       65   13.707    0.2109
---
```

example mangrove

Are there positive effects of two common sponge species and location on root growth of the mangrove tree, *Rhizophora mangle?*

```
sponge.lm <- lm(RootGrowthRate ~ Treatment + Location, data=sponge)
  summary(sponge.lm)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.19589    0.13777   1.422 0.159848
TreatmentFoam       0.35085    0.14356   2.444 0.017248 *
TreatmentHaliclona  0.47934    0.15025   3.190 0.002189 **
TreatmentTedania    0.59679    0.16502   3.617 0.000584 ***
Locationetb         0.24262    0.15069   1.610 0.112234
Locationlcn        -0.02891    0.15751  -0.184 0.854918
Locationlcs         0.01155    0.15201   0.076 0.939653
---
```

example mangrove

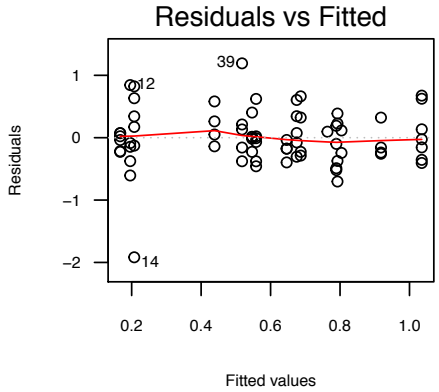
Are there positive effects of two common sponge species and location on root growth of the mangrove tree, *Rhizophora mangle?*

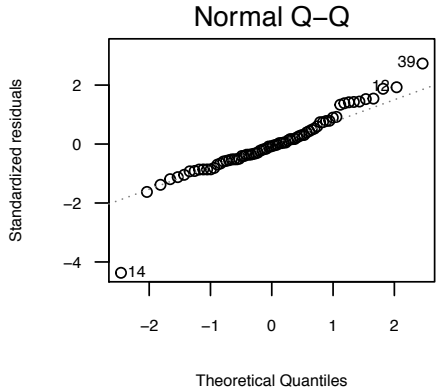
```
TukeyHSD(aov(sponge.lm)

       Tukey multiple comparisons of means
    95% family-wise confidence level

$Treatment
                       diff         lwr       upr     p adj
Foam-Control       0.3543571 -0.02395702 0.7326713 0.0743540
Haliclona-Control  0.4910924  0.09605067 0.8861342 0.0089071
Tedania-Control    0.6764286  0.25865046 1.0942067 0.0003723
Haliclona-Foam     0.1367353 -0.26270026 0.5361709 0.8034678
Tedania-Foam       0.3220714 -0.09986378 0.7440066 0.1939350
Tedania-Haliclona  0.1853361 -0.25166009 0.6223324 0.6796605

$Location
               diff        lwr       upr     p adj
etb-bbs   0.23085073 -0.1619963 0.6236978 0.4144956
lcn-bbs  -0.01714376 -0.4279931 0.3937055 0.9995184
lcs-bbs   0.01970554 -0.3785604 0.4179714 0.9991986
lcn-etb  -0.24799449 -0.6588438 0.1628548 0.3905194
lcs-etb  -0.21114519 -0.6094111 0.1871207 0.5052260
lcs-lcn   0.03684929 -0.3791844 0.4528830 0.9954699
```

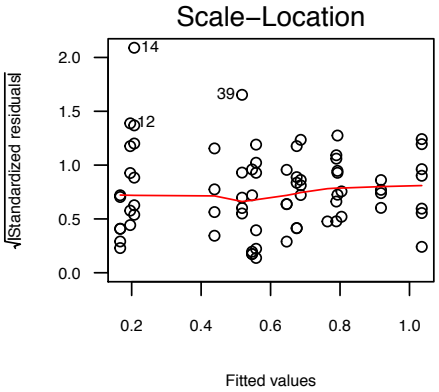
example

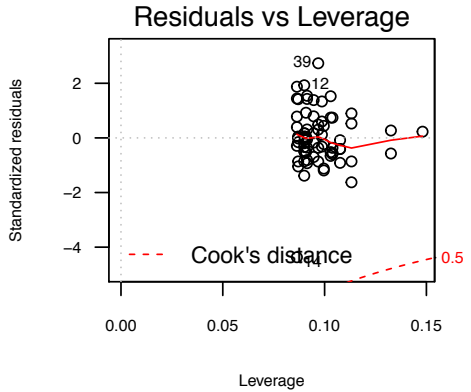
mangrove

```
plot(aov(RootGrowthRate ~
Treatment+Location))
```









Questions?