STAT 516 Lec 08

One- and two-way random effects models

Karl Gregory

2024-03-28

Fish net data from Dr. Longnecker's Notes

Strength of fish net material made from four randomly chosen machines.

	Strength Measurements: y_{ij}					
Machine	1	2	3	4	5	Mean:$\bar{y}_{i.}$
1	128	127	129	126	128	127.6
2	121	120	123	122	125	122.2
3	126	125	127	125	124	125.4
4	125	126	129	128	127	127.0

```
y <- c(128,127,129,126,128,121,120,123,122,125,
       126,125,127,125,124,125,126,129,128,127)
machine <- as.factor(c(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,4,4,4,4,4))
```

Is there significant machine-to-machine variability?

One-way random effects model

Suppose

$$Y_{ij} = \mu + A_i + \varepsilon_{ij}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, n_i$$

where

- Y_{ij} is the reponse of EU j in treatment group i.
- the A_i are independent Normal$(0, \sigma_A^2)$ rvs called random effects.
- the ε_{ij} are independent Normal$(0, \sigma_\varepsilon^2)$.
- the A_i and the ε_{ij} are independent of each other.
- μ is the overall mean
- we call σ_A^2 and σ_ε^2 variance components.

Until now we have studied fixed-effects models.

Assume a balanced design, i.e. $n_1 = \cdots = n_a = n$.

Goals in the one-way random effects model

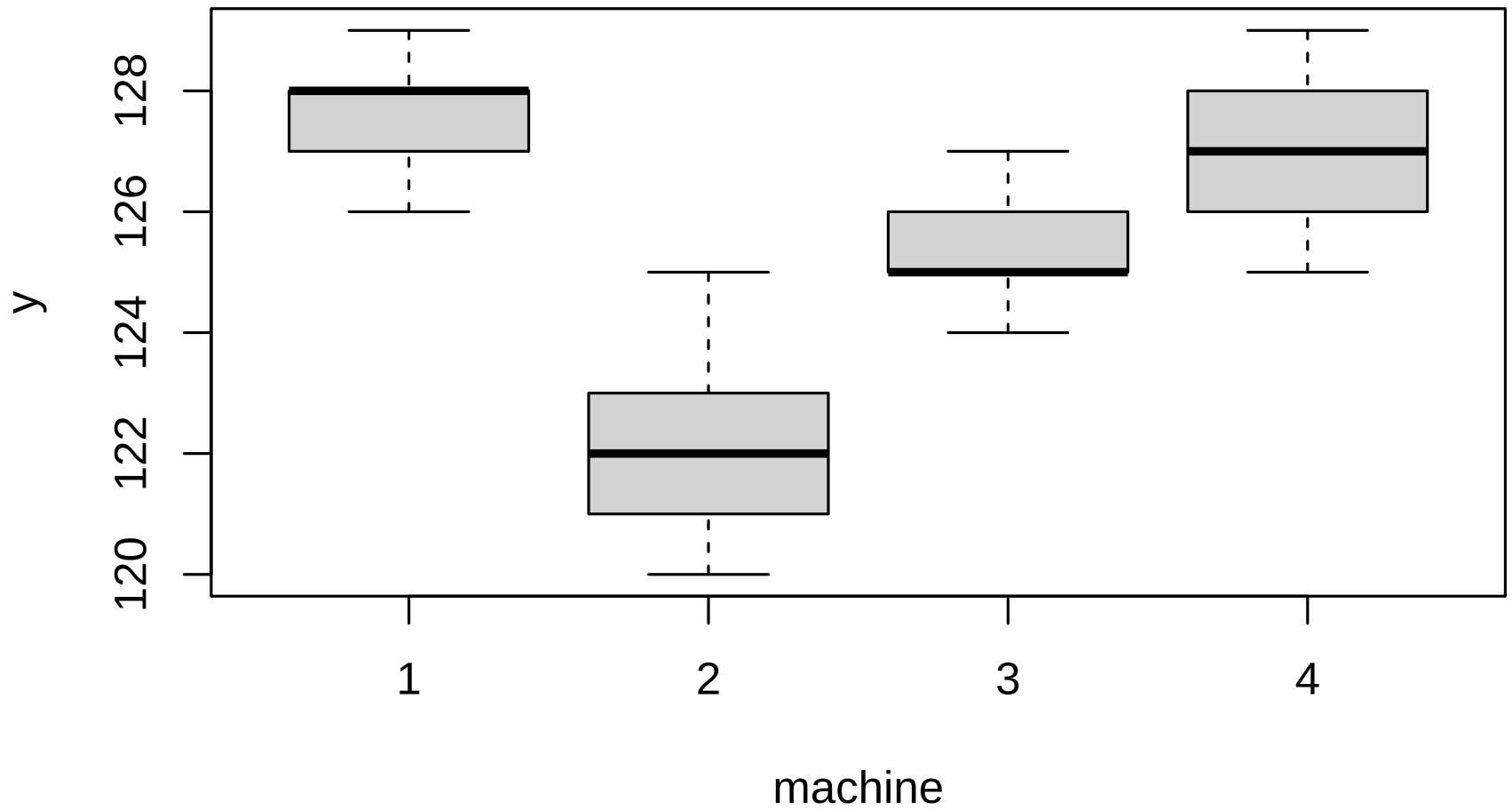
In the one-way random effects model we wish to

1. Visualize the data.
2. Decompose the variability in the Y_{ijk} into its sources.
3. Test H_0: $\sigma_A^2 = 0$ versus H_1: $\sigma_A^2 > 0$.
4. Estimate the variance components σ_A^2 and σ_ε^2.
5. Estimate the overall mean μ.
6. "Predict" the realized values of $A_1, \ldots, A_a$.
7. Check whether the model assumptions are satisfied.

Fish net data (cont)

Side-by-side boxplots are a natural choice if the n_i are not too small.

```
boxplot(y~machine)
```



Sums of squares for one-way random effects model

Sum of squares	Symbol	Formula
Total	$\mathrm{SS}_{\mathrm{Tot}}$	$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{\cdot\cdot})^2$
Treatment	$\mathrm{SS}_{\mathrm{Trt}}$	$\sum_{i=1}^{a} n(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$
Error	$\mathrm{SS}_{\mathrm{Error}}$	$\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i\cdot})^2$

We have the decomposition $\mathrm{SS}_{\mathrm{Tot}} = \mathrm{SS}_{\mathrm{Trt}} + \mathrm{SS}_{\mathrm{Error}}$.

ANOVA table for one-way random effects model

Obtain the MS values by dividing the SS values by the Df values.

Source	Df	SS	MS	F value
Treatment	$a-1$	$\mathrm{SS}_{\mathrm{Trt}}$	$\mathrm{MS}_{\mathrm{Trt}}$	$F_{\mathrm{stat}} = \mathrm{MS}_{\mathrm{Trt}} \,/\, \mathrm{MS}_{\mathrm{Error}}$
Error	$a(n-1)$	$\mathrm{SS}_{\mathrm{Error}}$	$\mathrm{MS}_{\mathrm{Error}}$	
Total	$an-1$	$\mathrm{SS}_{\mathrm{Tot}}$		

We reject H_0: $\sigma_A^2 = 0$ if $F_{\mathrm{stat}} > F_{a-1,a(n-1),\alpha}$.

The corresponding p-value is $P(F > F_{\mathrm{stat}})$, where $F \sim F_{a-1,a(n-1),\alpha}$.

Discuss: It can be shown that $\mathrm{SS}_{\mathrm{Trt}}$ and $\mathrm{SS}_{\mathrm{Error}}$ are independent and

$$\frac{\mathrm{SS}_{\mathrm{Trt}}}{n\sigma_A^2 + \sigma_\varepsilon^2} \sim \chi^2_{a-1} \quad \text{and} \quad \frac{\mathrm{SS}_{\mathrm{Error}}}{\sigma_\varepsilon^2} \sim \chi^2_{a(n-1)}.$$

Show that $F_{\mathrm{stat}} = \dfrac{\mathrm{MS}_{\mathrm{Trt}}}{\mathrm{MS}_{\mathrm{Error}}} \sim F_{a-1,a(n-1)}$ under H_0: $\sigma_A^2 = 0$.

Fish net data (cont)

We obtain the ANOVA table just as we did in the fixed-effects case.

```
lm_out <- lm(y~machine)
anova(lm_out)
```

```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
machine    3  87.75   29.25  13.296 0.0001301 ***
Residuals 16  35.20    2.20
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The small p-value indicates strong evidence against H_0: $\sigma_A^2 = 0$.

Expected mean squares in one-way random-effects model

In models with random effects, we like to track how the variance components σ_A^2 and σ_ε^2 affect the expected values of the mean squares.

Source	Df	Expected mean square
Treatment	$a-1$	$n\sigma_A^2+\sigma_\varepsilon^2$
Error	$a(n-1)$	σ_ε^2

Method of moments for estimating σ_A^2 and σ_ε^2

The method of moments (MoM) is one way to estimate the variance components.

1. Set the observed mean squares equal to the expected mean squares:

$$\mathrm{MS}_{\mathrm{Trt}} \overset{\mathrm{set}}{=} n\sigma_A^2 + \sigma_\varepsilon^2 \quad \text{and} \quad \mathrm{MS}_{\mathrm{Error}} \overset{\mathrm{set}}{=} \sigma_\varepsilon^2$$

2. Solve these equations for σ_A^2 and σ_ε^2, which gives

$$\dot{\sigma}_A^2 = \frac{\mathrm{MS}_{\mathrm{Trt}} - \mathrm{MS}_{\mathrm{Error}}}{n} \quad \text{and} \quad \dot{\sigma}_\varepsilon^2 = \mathrm{MS}_{\mathrm{Error}}\,.$$

MoM method is deprecated because it is possible to get $\dot{\sigma}_A^2 < 0$.

Fish net data (cont)

Compute the MoM estimators $\dot{\sigma}_A^2$ and $\dot{\sigma}_\varepsilon^2$ for the fish net data.

```
n <- 5
lm_out_anova <- anova(lm_out)
MSTrt <- lm_out_anova$`Mean Sq`[1]
MSE <- lm_out_anova$`Mean Sq`[2]
se2_dot <- MSE
sA2_dot <- (MSTrt - MSE)/n
```

We obtain $\dot{\sigma}_A^2 = 5.41$ and $\dot{\sigma}_\varepsilon^2 = 2.2$.

REML for estimating the variance components

- REML stands for restricted maximum likelihood.
- It is a recipe (beyond course) for estimating variance components.
- In the one-way random-effects model it results in

$$\hat{\sigma}_A^2 = \begin{cases} \dot{\sigma}_A^2 & \text{if } \dot{\sigma}_A^2 \geq 0 \\ 0 & \text{if } \dot{\sigma}_A^2 < 0 \end{cases} \quad \text{and} \quad \hat{\sigma}_\varepsilon^2 = \begin{cases} \dot{\sigma}_\varepsilon^2 & \text{if } \dot{\sigma}_A^2 \geq 0 \\ \dfrac{\mathrm{SS}_{\mathrm{Tot}}}{an-1} & \text{if } \dot{\sigma}_A^2 < 0 \end{cases}$$

- Will always return a nonnegative estimate of σ_A^2.
- Same as MoM estimator if $\dot{\sigma}_A^2 > 0$.
- The "state of the art" is to use REML estimation.

Fish net data (cont)

Use the R package `lme4` to compute the REML estimates of σ_A^2 and σ_ε^2 on the fish net data.

```
library(lme4) # may need to run install.packages("Matrix") for this to work
lmer_out <- lmer(y ~ 1 + (1 | machine))
summary(lmer_out)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ 1 + (1 | machine)

REML criterion at convergence: 79.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.6531 -0.6884 -0.1019  0.4846  1.7179

Random effects:
 Groups   Name        Variance Std.Dev.
 machine  (Intercept) 5.41     2.326
 Residual             2.20     1.483
Number of obs: 20, groups:  machine, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)  125.550      1.209   103.8
```

Estimating μ and predicting $A_1, \ldots, A_a$.

- We estimate the overall mean μ with $\bar{Y}_{\cdot\cdot}$.
- We also wish to guess at the realized values of $A_1, \ldots, A_a$.
- Call this *prediction* instead of estimation, because $A_1, \ldots, A_a$ are random variables rather than parameters with fixed values.
- A recipe (beyond course) for the best linear unbiased predictor yields

$$\hat{A}_i = \left(\frac{n\hat{\sigma}_A^2}{n\hat{\sigma}_A^2 + \hat{\sigma}_\varepsilon^2} \right) (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) \quad \text{for } i = 1, \ldots, a.$$

Fish net data (cont)

Obtain the predicted values of A_1, A_2, A_3, and A_4 for the fish net data.

```
yi.bar <- aggregate(y, by = list(machine = machine), mean)$x
y..bar <- mean(y)
# use MoMs, since sA2_dot > 0
Ahat <- (n * sA2_dot)/(n * sA2_dot + se2_dot) * (yi.bar - y..bar)
Ahat
```

```
[1]  1.8958120 -3.0980342 -0.1387179  1.3409402
```

Can use `predict()` on the `lmer()` output to obtain the $\hat{A}_i$ values:

```
predict(lmer_out,random.only = TRUE) # gives estimates of the A_i for each obs
```

```
         1          2          3          4          5          6          7
 1.8958120  1.8958120  1.8958120  1.8958120  1.8958120 -3.0980342 -3.0980342
         8          9         10         11         12         13         14
-3.0980342 -3.0980342 -3.0980342 -0.1387179 -0.1387179 -0.1387179 -0.1387179
        15         16         17         18         19         20
-0.1387179  1.3409402  1.3409402  1.3409402  1.3409402  1.3409402
```

Fitted values and residuals in the random effects model

- Build fitted values from $\hat{\mu}$ and the predictions $\hat{A}_1, \ldots, \hat{A}_a$ as

$$\hat{Y}_{ij} = \hat{\mu} + \hat{A}_i, \quad \text{for } i = 1, \ldots, a, \quad j = 1, \ldots, n.$$

- Our residuals then become

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij}, \quad \text{for } i = 1, \ldots, a, \quad j = 1, \ldots, n.$$

Use `predict()` on the `lmer()` output to obtain the $\hat{Y}_{ij} = \hat{\mu} + \hat{A}_i$.

```
predict(lmer_out) # gives estimates of mu + A_i
```

```
       1        2        3        4        5        6        7        8
127.4458 127.4458 127.4458 127.4458 127.4458 122.4520 122.4520 122.4520
       9       10       11       12       13       14       15       16
122.4520 122.4520 125.4113 125.4113 125.4113 125.4113 125.4113 126.8909
      17       18       19       20
126.8909 126.8909 126.8909 126.8909
```

Checking model assumptions

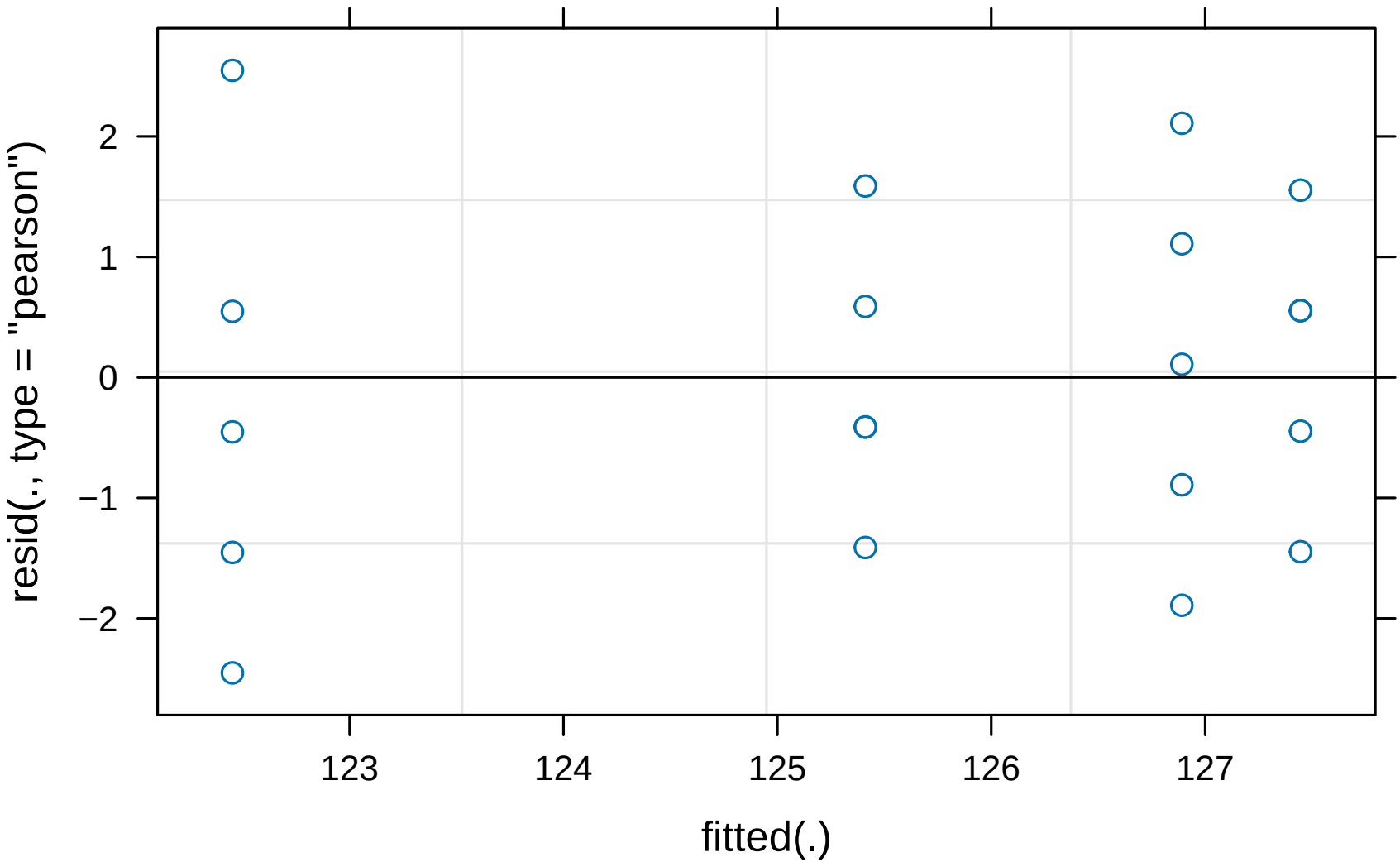
Validity of the methods in these slides depends on these assumptions:

1. The responses are normally distributed around the treatment means.
   (Check QQ plot of residuals)
2. The treatment means are normally distributed.
   (The number of groups a is often too small for this to be checked)
3. The response has the same variance in all treatment groups.
   (Check residuals vs fitted values plot)
4. The response values in each treatment group are independent.
   (No way to check; must trust experimental design)

Fish net data (cont)

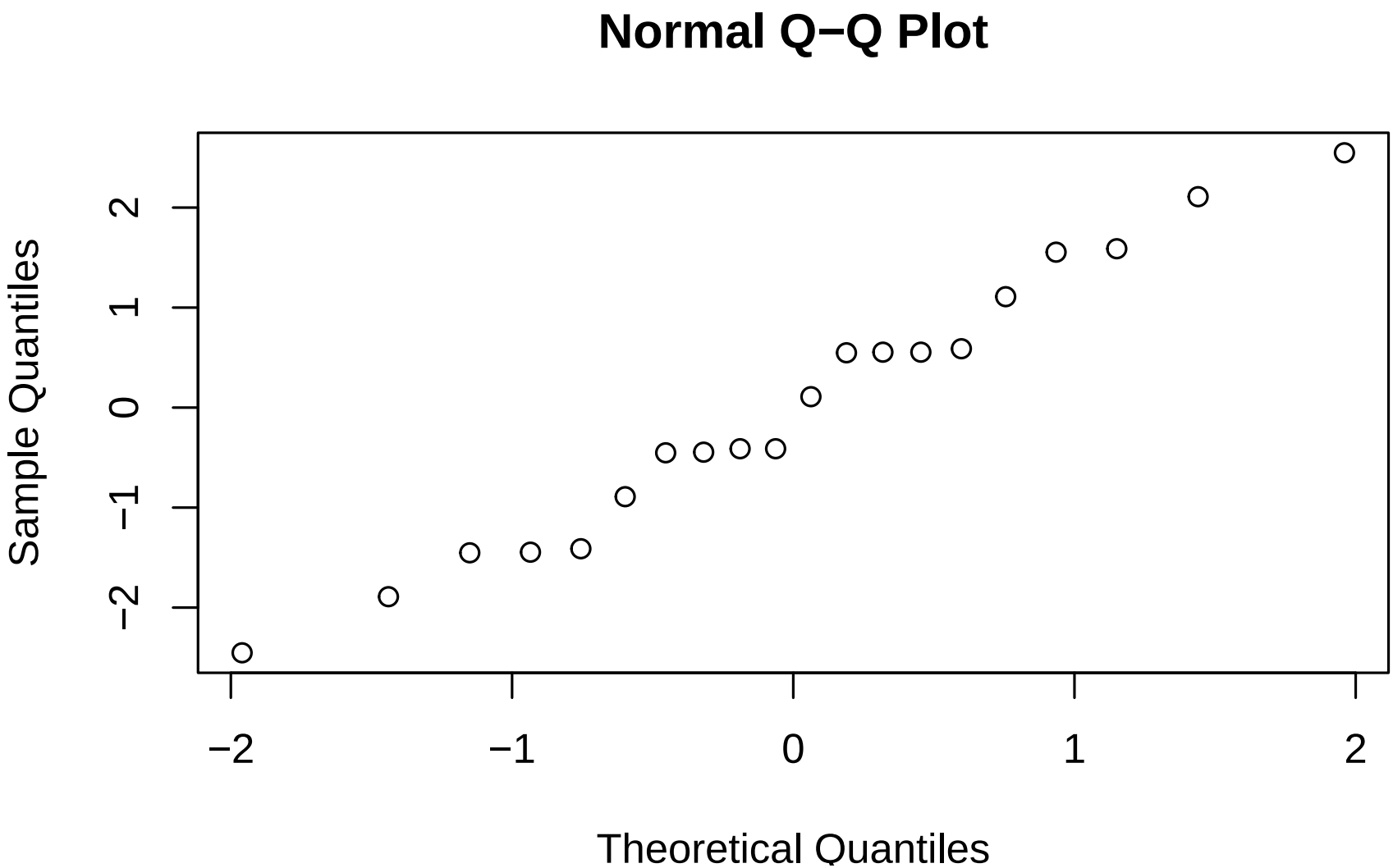
Check the residuals vs fitted values plot: Use `plot()` on `lmer()` output.

```
plot(lmer_out)
```



Check the normal q-q plot of the residuals:

```
yhat <- predict(lmer_out)
ehat <- y - yhat
qqnorm(ehat)
```



Triglyceride level data from Kuehl (2000)

Triglyceride levels in serum samples measured with four machines on four days.

	Machine			
Day	1	2	3	4
1	142.3, 144.0	148.6, 146.9	142.9, 147.4	133.8, 133.2
2	134.9, 146.3	145.2, 146.3	125.9, 127.6	108.9, 107.5
3	148.6, 156.5	148.6, 153.1	135.5, 138.9	132.1, 149.7
4	152.0, 151.4	149.7, 152.0	142.9, 142.3	141.7, 141.2

Source: Dr. J. Anderson, Beckman Instruments, Inc.

```
tg <- c(142.3,144.0,148.6,146.9,142.9,147.4,133.8,133.2,
        134.9,146.3,145.2,146.3,125.9,127.6,108.9,107.5,
        148.6,156.5,148.6,153.1,135.5,138.9,132.1,149.7,
        152.0,151.4,149.7,152.0,142.9,142.3,141.7,141.2)
days <- as.factor(c(rep(1,8),rep(2,8),rep(3,8),rep(4,8)))
machine <- as.factor(c(1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4,
                       1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4))
```

Machines randomly sampled. Regard days as randomly sampled.

Two-way random effects model

Assume

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + \varepsilon_{ijk}$$

for $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $n = 1, \ldots, n$, where

- $A_i \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma_A^2)$
- $B_j \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma_B^2)$
- $(AB)_{ij} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma_{AB}^2)$
- $\varepsilon_{ijk} \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma_\varepsilon^2)$
- the A_i, B_j, and ε_{ijk} are all independent.
- μ is the overall mean.
- we call σ_A^2, σ_B^2, σ_{AB}^2, and σ_ε^2 variance components.

Still assume a balanced design, that is $n_{ij} = n$ for all i, j.

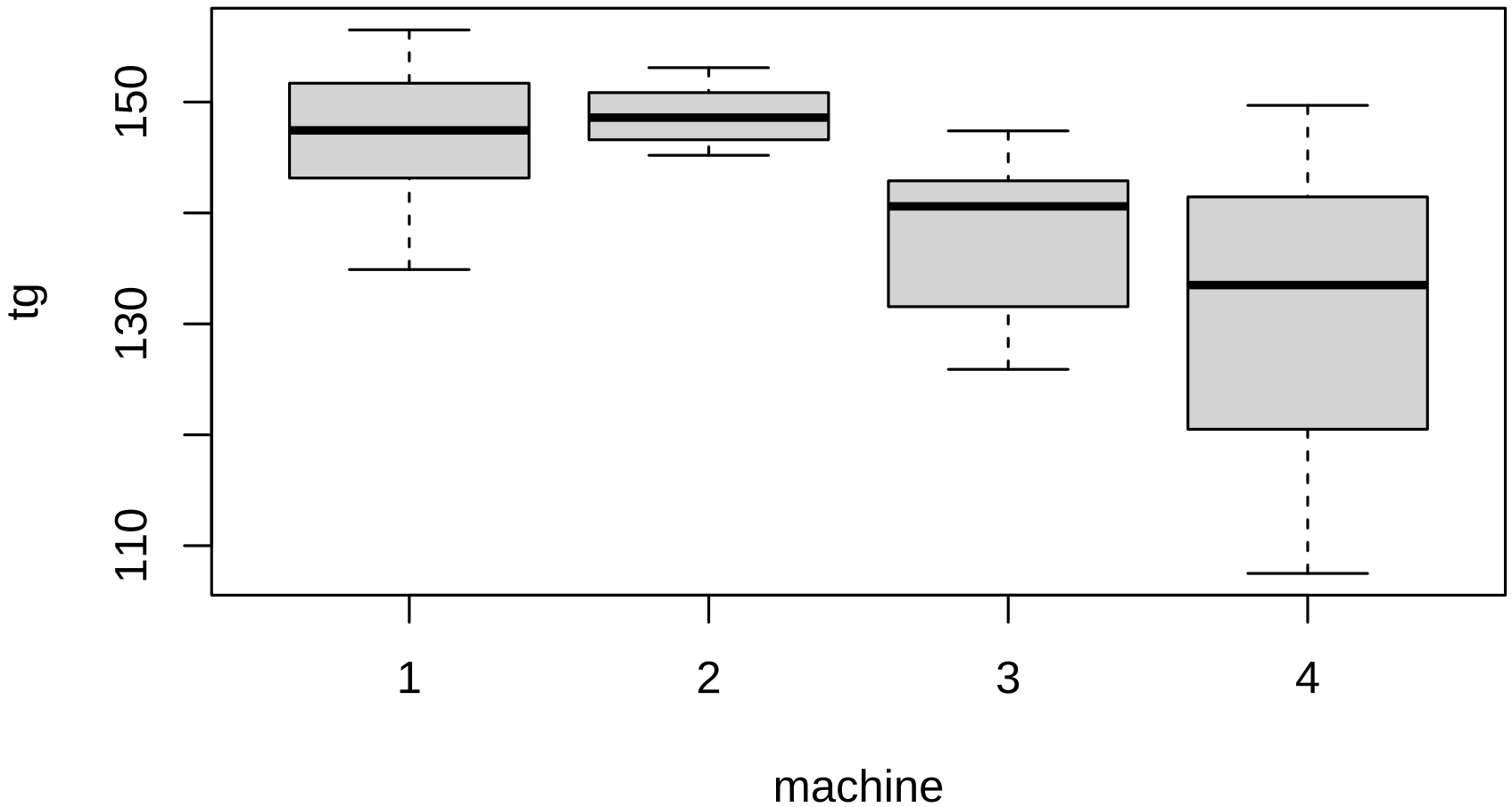
Goals in the two-way random effects model

In the two-way random effects model we wish to

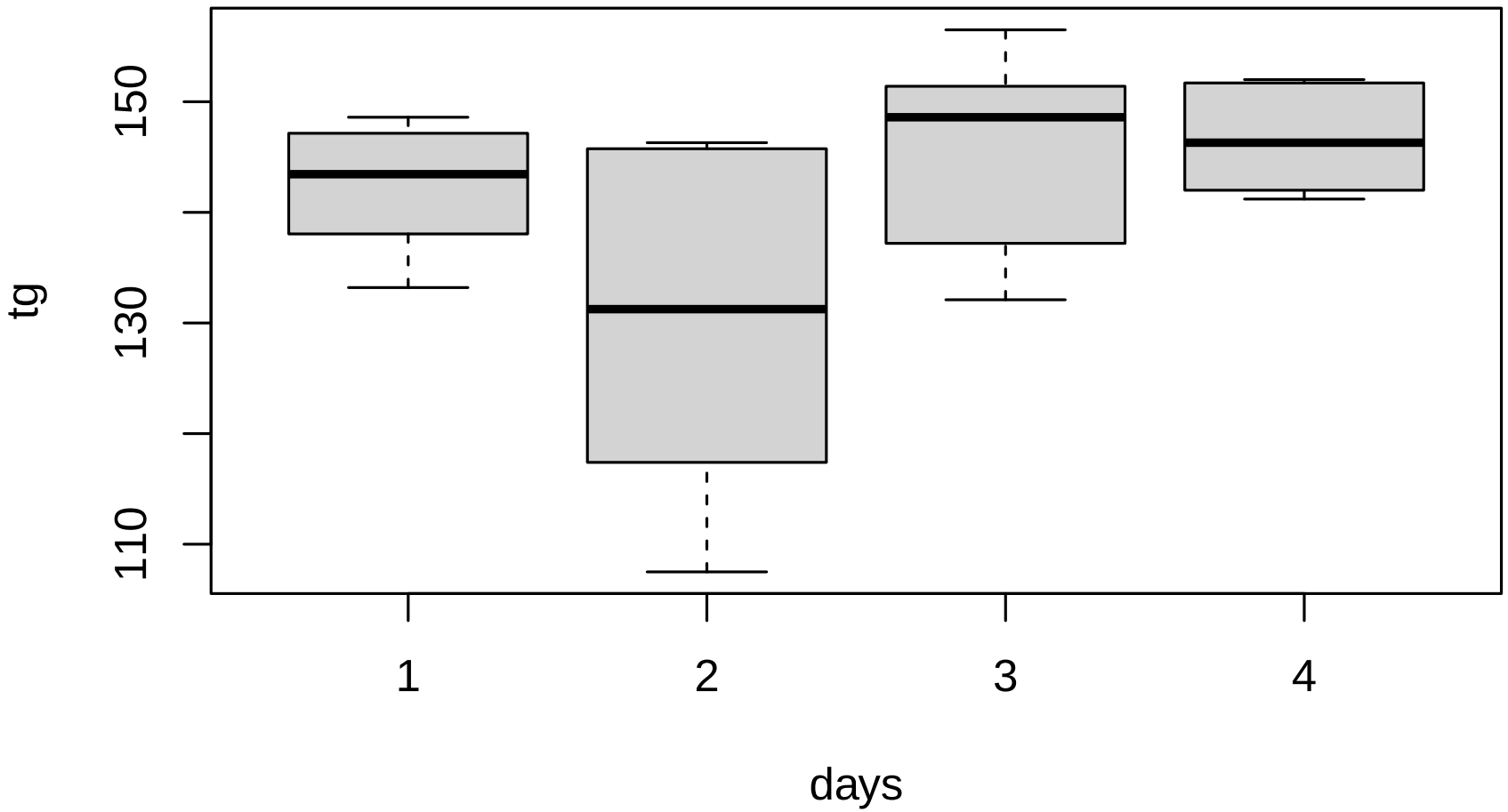
1. Visualize the data.
2. Decompose the variability in the Y_{ijk} into its sources.
3. Test whether σ^2_A, σ^2_B, and σ^2_{AB} are equal to zero.
4. Estimate the variance components σ^2_A, σ^2_B, σ^2_{AB}, and σ^2_ε.
5. Estimate the overall mean μ.
6. "Predict" the realized values of the random effects.
7. Check whether the model assumptions are satisfied.

Triglyceride level data (cont)

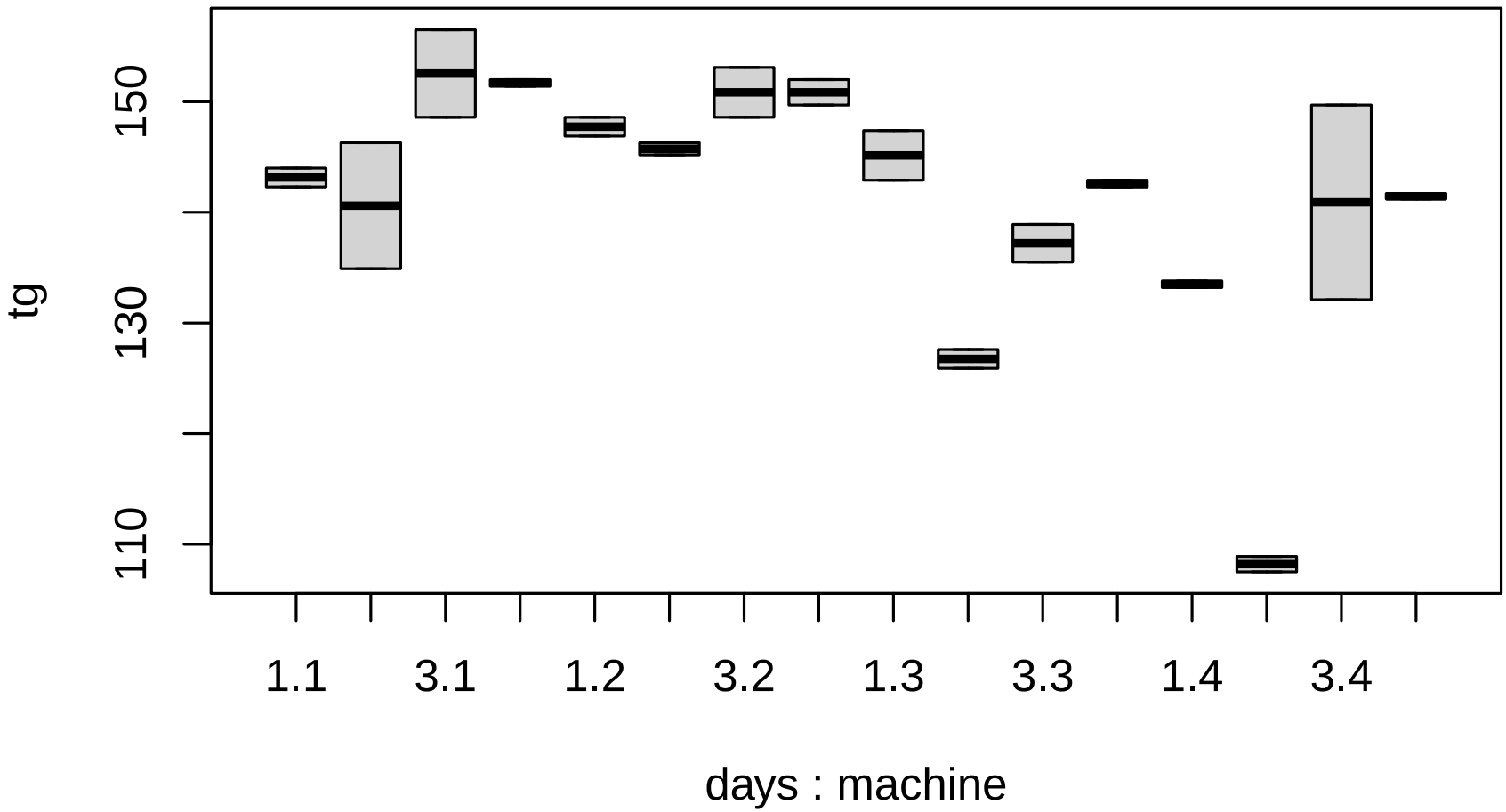
```
boxplot(tg~machine)
```



```
boxplot(tg~days)
```



```
boxplot(tg~days + machine)
```



Sums of squares for the two-way random effects model

Sum of squares	Symbol	Formula
Total	$\mathrm{SS}_{\mathrm{Tot}}$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\bar{Y}_{...})^2$
A	SS_{A}	$nb\sum_{i=1}^{a}(\bar{Y}_{i..}-\bar{Y}_{...})^2$
B	SS_{B}	$na\sum_{j=1}^{b}(\bar{Y}_{.j.}-\bar{Y}_{...})^2$
AB	$\mathrm{SS}_{\mathrm{AB}}$	$n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_{ij.}-(\bar{Y}_{i..}+\bar{Y}_{.j.}-\bar{Y}_{...})^2$
Error	$\mathrm{SS}_{\mathrm{Error}}$	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\bar{Y}_{ij.})^2$

We have the decomposition $\mathrm{SS}_{\mathrm{Tot}} = \mathrm{SS}_{\mathrm{A}} + \mathrm{SS}_{\mathrm{B}} + \mathrm{SS}_{\mathrm{AB}} + \mathrm{SS}_{\mathrm{Error}}$.

Full ANOVA table for balanced two-way factorial design

Obtain the MS values by dividing the SS values by the Df values.

Source	Df	SS	MS	F value
A	$a-1$	SS_{A}	MS_{A}	$F_{\mathrm{A}} = \mathrm{MS}_{\mathrm{A}} / \mathrm{MS}_{\mathrm{AB}}$
B	$b-1$	SS_{B}	MS_{B}	$F_{\mathrm{B}} = \mathrm{MS}_{\mathrm{B}} / \mathrm{MS}_{\mathrm{AB}}$
AB	$(a-1)(b-1)$	$\mathrm{SS}_{\mathrm{AB}}$	$\mathrm{MS}_{\mathrm{AB}}$	$F_{\mathrm{AB}} = \mathrm{MS}_{\mathrm{AB}} / \mathrm{MS}_{\mathrm{Error}}$
Error	$ab(n-1)$	$\mathrm{SS}_{\mathrm{Error}}$	$\mathrm{MS}_{\mathrm{Error}}$	
Total	$abn-1$	$\mathrm{SS}_{\mathrm{Tot}}$		

1. Reject H_0: $\sigma_A^2 = 0$ if $F_{\mathrm{A}} > F_{a-1,(a-1)(b-1),\alpha}$.
2. Reject H_0: $\sigma_B^2 = 0$ if $F_{\mathrm{B}} > F_{b-1,(a-1)(b-1),\alpha}$.
3. Reject H_0: $\sigma_{AB}^2 = 0$ if $F_{\mathrm{AB}} > F_{(a-1)(b-1),ab(n-1),\alpha}$.

Note F_{A} and F_{B} are different from their fixed-effects counterparts!

Expected mean squares in two-way random-effects model

We again tabulate the expected values of the mean squares:

Source	Df	Expected mean square
A	$a-1$	$nb\sigma_A^2 + n\sigma_{AB}^2 + \sigma_\varepsilon^2$
B	$b-1$	$na\sigma_B^2 + n\sigma_{AB}^2 + \sigma_\varepsilon^2$
AB	$(a-1)(b-1)$	$n\sigma_{AB}^2 + \sigma_\varepsilon^2$
Error	$a(n-1)$	σ_ε^2

Discuss: Make F-stat by dividing by the MS having the same expectation under H_0.

Triglyceride level data (cont)

Obtain p-values for testing whether the variance components are zero.

The `anova()` function on the `lm()` output only gives the correct F statistic and p-value for the interaction.

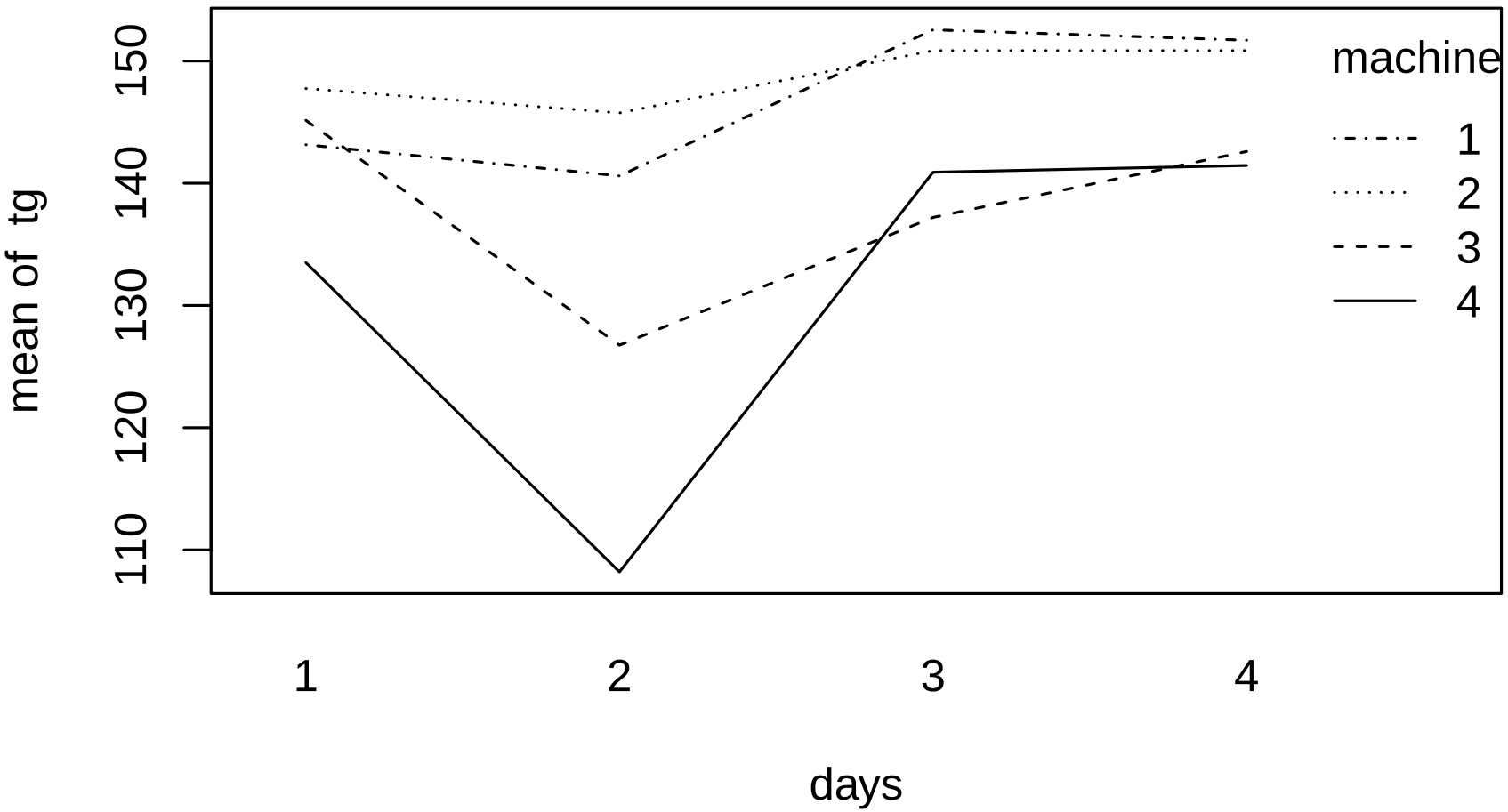
```
anova_out <- anova(lm(tg ~ days + machine + days:machine))
anova_out
```

```
Analysis of Variance Table

Response: tg
             Df  Sum Sq Mean Sq F value    Pr(>F)
days          3 1334.46  444.82 24.8569 2.907e-06 ***
machine       3 1647.28  549.09 30.6836 7.192e-07 ***
days:machine  9  786.04   87.34  4.8805  0.002936 **
Residuals    16  286.33   17.90
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the p-value for H_0: $\sigma^2_{AB} = 0$, we conclude there is interaction between the day and the machine. So machine performance does not appear to be consistent across days.

```
interaction.plot(days,machine,tg)
```



```r
a <- 4
b <- 4
n <- 2
MSA <- anova_out$`Mean Sq`[1]
MSB <- anova_out$`Mean Sq`[2]
MSAB <- anova_out$`Mean Sq`[3]
FA <- MSA / MSAB
FB <- MSB / MSAB
pA <- 1 - pf(FA,a - 1,(a-1)*(b-1))
pB <- 1 - pf(FB,b - 1,(a-1)*(b-1))
```

For H_0: $\sigma_A^2 = 0$ we obtain a p-value of 0.0248.
For H_0: $\sigma_B^2 = 0$ we obtain a p-value of 0.0137.

There is significant machine-to-machine and day-to-day variation.

Estimation of σ^2_A, σ^2_B, σ^2_{AB}, and σ^2_ε

- Could use the MoM based on the table of expected mean squares.
- MoM can result is negative estimates of σ^2_A, σ^2_B, or σ^2_{AB}.
- REML estimation is prescribed.
- Estimate μ with the overall data mean $\bar{Y}_{\cdots}$.

Triglyceride level data (cont)

Obtain the REML estimators of σ^2_A, σ^2_B, σ^2_{AB}, and σ^2_ε.

Use `summary()` on the `lmer()` output.

```
lmer_out <- lmer(tg ~ 1 + (1|days) + (1|machine) + (1|days:machine))
summary(lmer_out)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: tg ~ 1 + (1 | days) + (1 | machine) + (1 | days:machine)

REML criterion at convergence: 215

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.84283 -0.35581  0.03485  0.20700  2.31766

Random effects:
 Groups       Name        Variance Std.Dev.
 days:machine (Intercept) 34.72    5.892
 machine      (Intercept) 57.72    7.597
 days         (Intercept) 44.69    6.685
 Residual                 17.90    4.230
Number of obs: 32, groups:  days:machine, 16; machine, 4; days, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)  141.184      5.323   26.52
```

Prediction of realized values of random effects

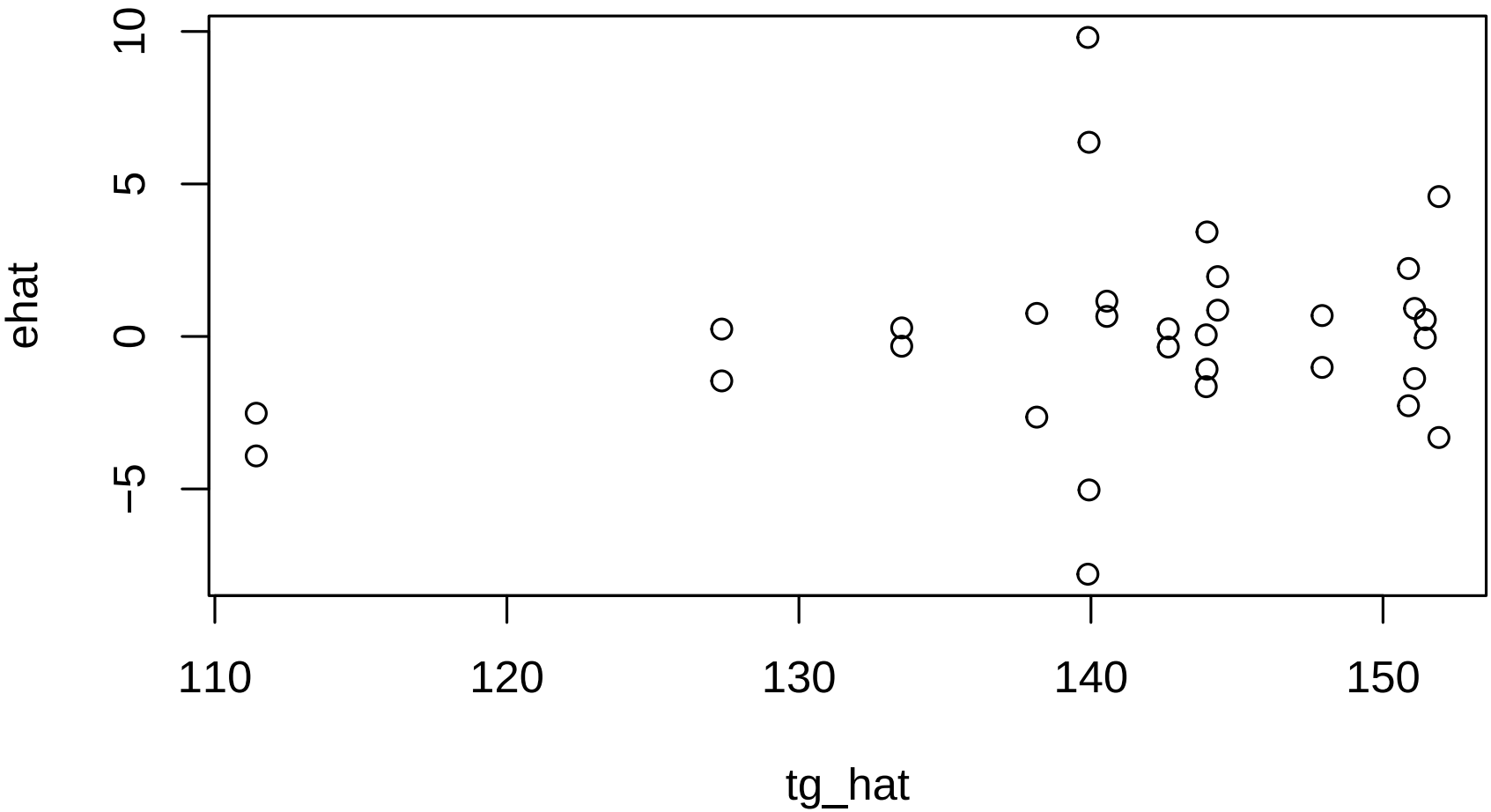
- We wish to guess at the values of A_i, B_j, and $(AB)_{ij}$ for all i, j.
- The formulas for the best linear unbiased predictors are complicated.
- The fitted values for the model are

$$\hat{Y}_{ijk} = \hat{\mu} + \hat{A}_i + \hat{B}_j + \widehat{(AB)}_{ij}, \qquad \text{for all } i, j, k$$

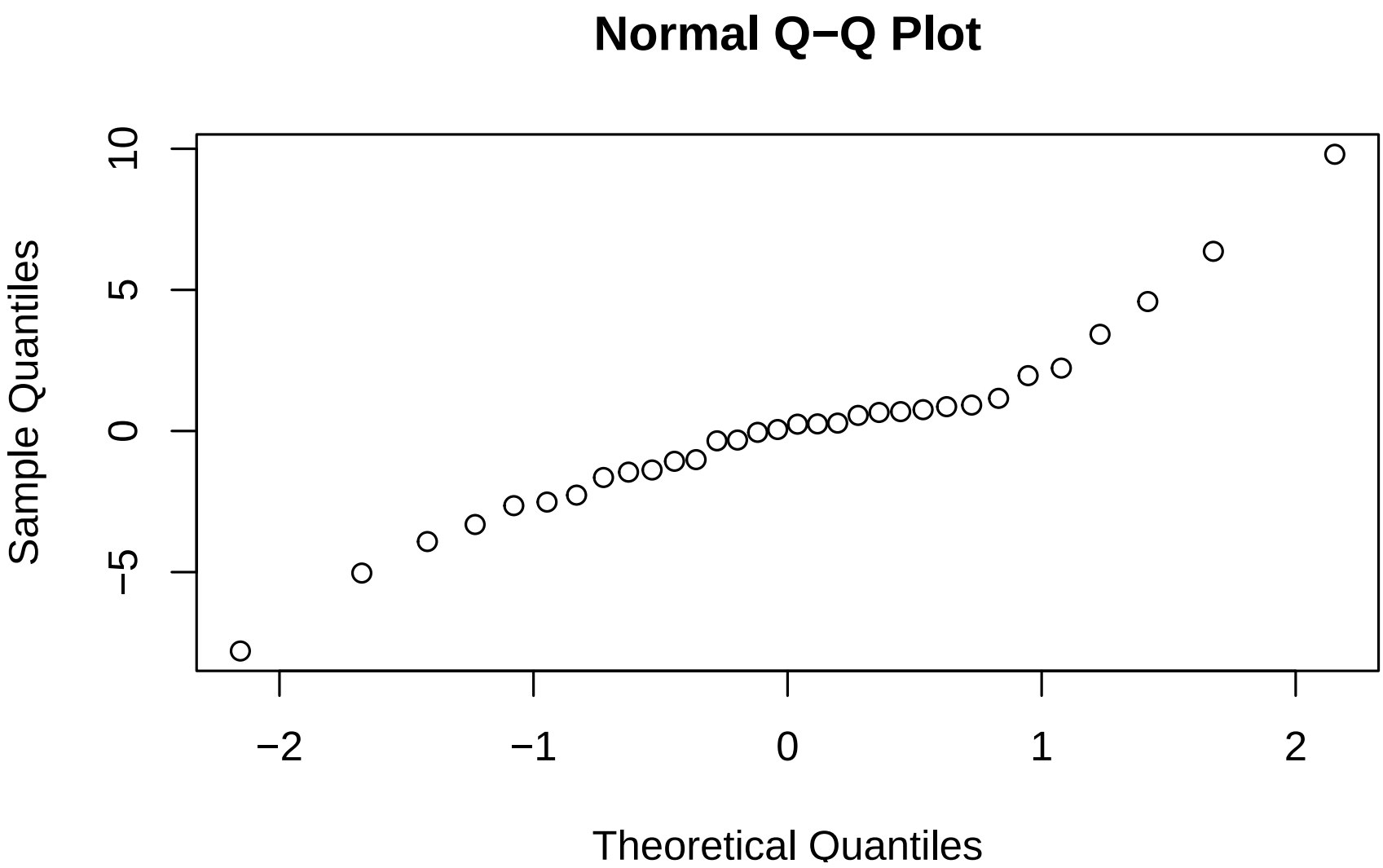
- Use `predict()` on the `lmer()` output to obtain these fitted values.
- The residuals are then $Y_{ijk} - \hat{Y}_{ijk}$ for all i, j, k.
- Can then check diagnostic plots.

Triglyceride level data (cont)

```
tg_hat <- predict(lmer_out)
ehat <- tg - tg_hat
plot(ehat ~ tg_hat)
```



```
qqnorm(ehat)
```



Both plots show some problems; interpret our results with caution.

References

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