Class 13 lab - no code

John Kubale

2024-11-26

Data Set-Up

- We will `warpbreaks` data in R.
- Description: The number of breaks in yarn during Weaving
  - This data set gives the number of warp breaks per loom, where a loom corresponds to a fixed length of yarn.
  - The type of wool and the level of weaving tension were randomly assigned to examine warp breaks.
- Format: A data frame with 54 observations on 3 variables.
  - `break`: numeric, The number of breaks
  - `wool`: factor, The type of wool (A or B)
  - `tension`: factor, The level of tension (L, M, H)
- See https://rdrr.io/r/datasets/warpbreaks.html for detail of `warpbreak`

1. What design structure do you see?

- It is a balance design.

```
data(warpbreaks)
summary(warpbreaks)
```

```
##      breaks      wool   tension
##  Min.   :10.00   A:27   L:18
##  1st Qu.:18.25   B:27   M:18
##  Median :26.00          H:18
##  Mean   :28.15
##  3rd Qu.:34.00
##  Max.   :70.00
```

```
warpbreaks%>%
  tabyl(wool,tension) # tabyl() is from the janitor package
```

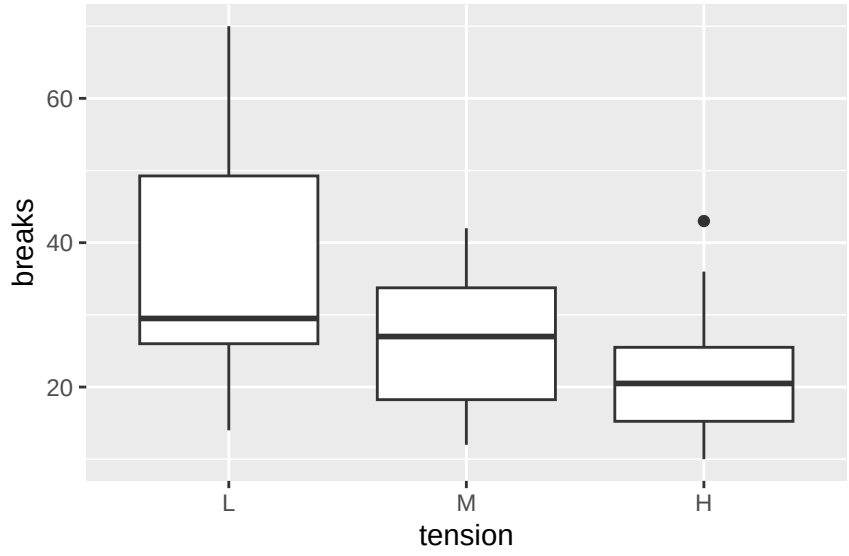
```
##  wool L M H
##     A 9 9 9
##     B 9 9 9
```

2. We will focus on `tension` as an experimental factor for `breaks` in this lab session.

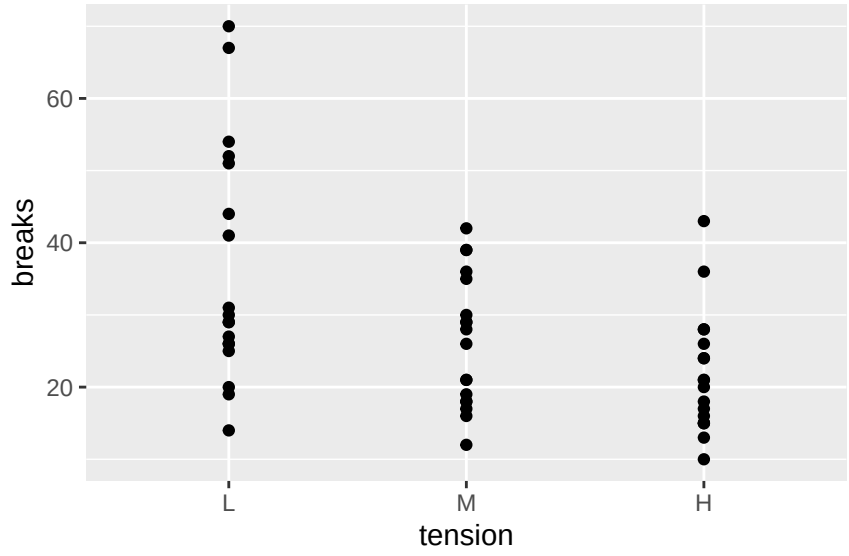
2.A. Examine the pattern of `break` by `tension` using plots

- Boxplot relatively linear association. The group means scatter plot included in the center, and the violin plot shows more clearly where data lies. The breaks + 3, if lower the means will go down.

```
ggplot(warpbreaks, aes(x=tension, y=breaks))+geom_boxplot()
```



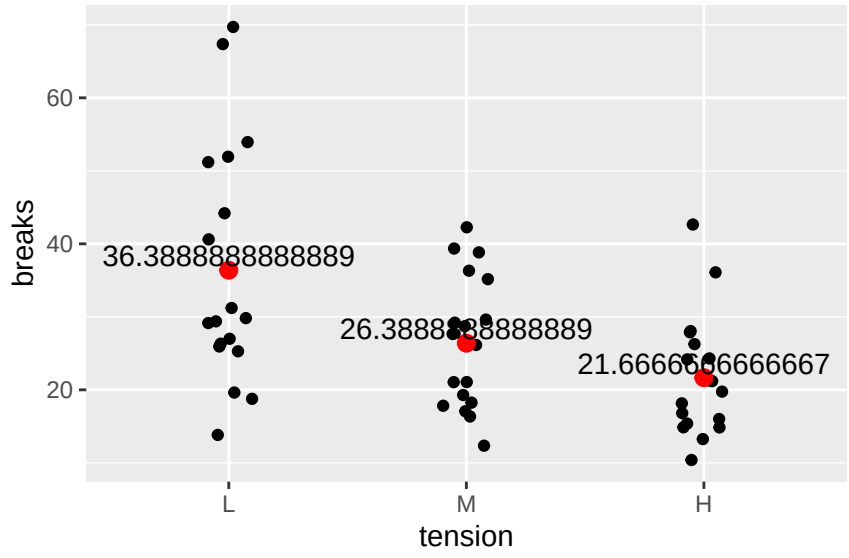

```
ggplot(warpbreaks, aes(x=tension, y=breaks))+geom_point()
```




```
(means <- aggregate(breaks~tension,warpbreaks, mean)) # aggregate works similarly to gr
```

```
##   tension   breaks
## 1       L 36.38889
## 2       M 26.38889
## 3       H 21.66667
```

```
ggplot(warpbreaks,aes(x=tension, y=breaks))+
  geom_point(position=position_jitter(width=0.1))+
  stat_summary(fun="mean",colour = "red")+
  geom_text(data = means, aes(label = breaks, y = breaks + 2))
```



```
# way too many digits after the decimal point so will round to 2
ggplot(warpbreaks,aes(x=tension, y=breaks))+
  geom_violin()+
  stat_summary(fun="mean", colour = "red")+
  geom_text(data = means, aes(label = round(breaks,2), y = breaks + 3),color="red")
```

What is the stat_summary() function doing in each plot? What happens if you change the number added to breaks in the last line of code?

2.B. One-way ANOVA of `breaks`

```
summary(aov(breaks~tension,warpbreaks))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## tension      2   2034  1017.1   7.206 0.00175 **
## Residuals   51   7199   141.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(breaks~tension,warpbreaks))
```

```
## Analysis of Variance Table
##
## Response: breaks
##           Df Sum Sq Mean Sq F value   Pr(>F)
## tension    2 2034.3 1017.13  7.2061 0.001753 **
```

```
## Residuals 51 7198.6  141.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(breaks~tension,warpbreaks))
```

```
##
## Call:
## lm(formula = breaks ~ tension, data = warpbreaks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.389  -8.139  -2.667   6.333  33.611
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.39       2.80  12.995  < 2e-16 ***
## tensionM      -10.00       3.96  -2.525 0.014717 *
## tensionH      -14.72       3.96  -3.718 0.000501 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.88 on 51 degrees of freedom
## Multiple R-squared:  0.2203, Adjusted R-squared:  0.1898
## F-statistic: 7.206 on 2 and 51 DF,  p-value: 0.001753
```

How do anova(), summary(aov()), and summary(lm()) compare? - aov() is equivalent to the OLS simple model.

2.C. Tukey's honest significance difference test

```
TukeyHSD(aov(breaks~tension,warpbreaks))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = breaks ~ tension, data = warpbreaks)
##
## $tension
##           diff       lwr        upr     p adj
## M-L -10.000000 -19.55982 -0.4401756 0.0384598
## H-L -14.722222 -24.28205 -5.1623978 0.0014315
## H-M  -4.722222 -14.28205  4.8376022 0.4630831
```

```
t.test(breaks~tension,subset(warpbreaks,tension!="H"))$p.value
```

```
## [1] 0.03252481
```

```
t.test(breaks~tension,subset(warpbreaks,tension!="M"))$p.value
```

```
## [1] 0.002326794
```

```
t.test(breaks~tension,subset(warpbreaks,tension!="L"))$p.value
```

```
## [1] 0.1145571
```

How do the p values from TukeyHSD() compare to those from t.test()? - We can see the CI for each difference and adjusted p-value because we have multiple comparisons.

2.D. Diagnostics for the error assumptions

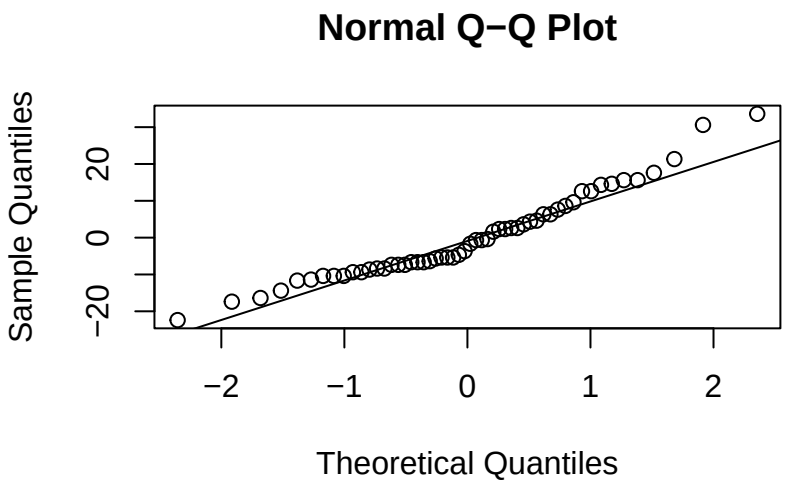
```
one<-lm(breaks~tension,warpbreaks)
summary(one)
```

```
##
## Call:
## lm(formula = breaks ~ tension, data = warpbreaks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.389  -8.139  -2.667   6.333  33.611
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.39       2.80  12.995  < 2e-16 ***
## tensionM      -10.00       3.96  -2.525 0.014717 *
## tensionH      -14.72       3.96  -3.718 0.000501 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.88 on 51 degrees of freedom
## Multiple R-squared:  0.2203, Adjusted R-squared:  0.1898
## F-statistic: 7.206 on 2 and 51 DF,  p-value: 0.001753
```

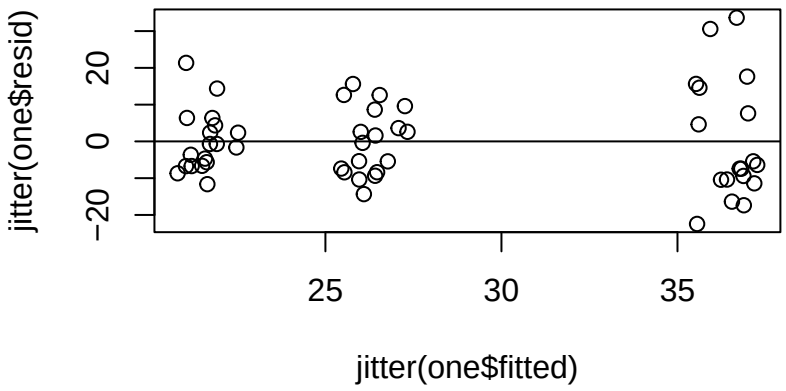
```r
anova(one)
```

```
## Analysis of Variance Table
##
## Response: breaks
##           Df Sum Sq Mean Sq F value   Pr(>F)
## tension    2 2034.3 1017.13  7.2061 0.001753 **
## Residuals 51 7198.6  141.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

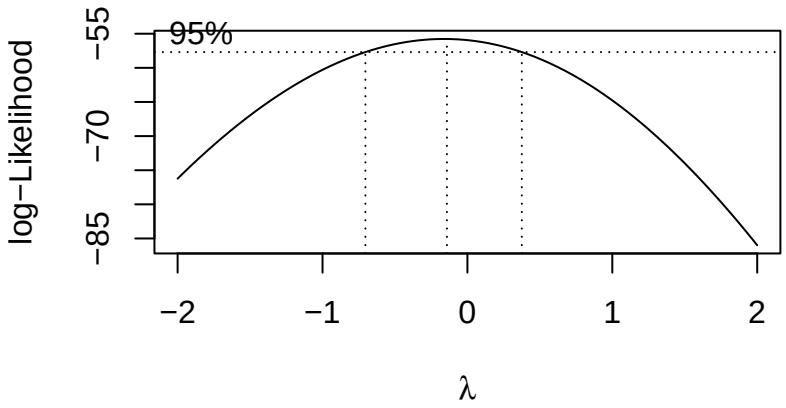
```r
qqnorm(one$resid);qqline(one$resid)
```




```r
plot(jitter(one$fitted),jitter(one$resid));abline(h=0)
```




```
boxcox(lm(breaks~tension,warpbreaks))
```



What does the jitter function do? - spreads the data so they do not overlap.

What (if any) transformation of the outcome is suggested? - Transformation suggested log function.

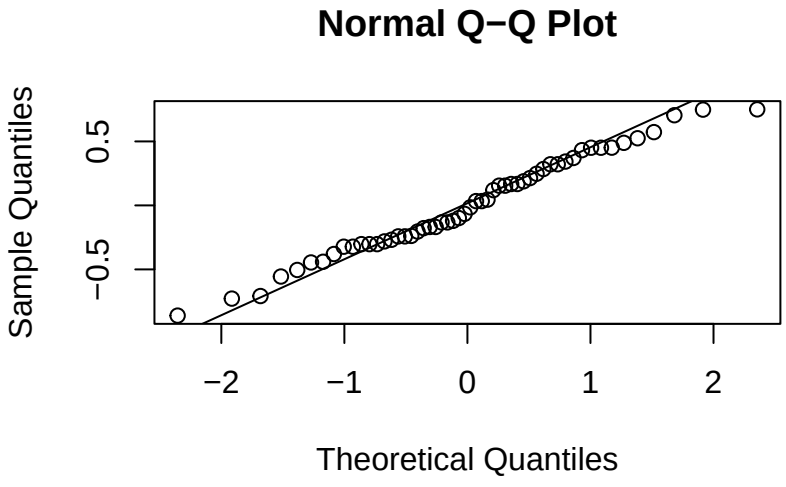
2.E. One-way ANOVA of `breaks` with transformed `break`

Fit a linear regression model with breaks as a function of tension. Transform breaks according to what you found in the previous question. Reassess OLS assumptions and conduct pairwise comparisons using TukeyHSD().

```r
one_t <- lm(log(breaks)~tension,warpbreaks)
summary(one_t)
```

```
##
## Call:
## lm(formula = log(breaks) ~ tension, data = warpbreaks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86110 -0.27811 -0.04066  0.31199  0.75031
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.5002     0.0930  37.637  < 2e-16 ***
## tensionM     -0.2871     0.1315  -2.183 0.033654 *
## tensionH     -0.4893     0.1315  -3.720 0.000497 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3946 on 51 degrees of freedom
## Multiple R-squared:  0.2151, Adjusted R-squared:  0.1843
## F-statistic: 6.989 on 2 and 51 DF,  p-value: 0.002077
```

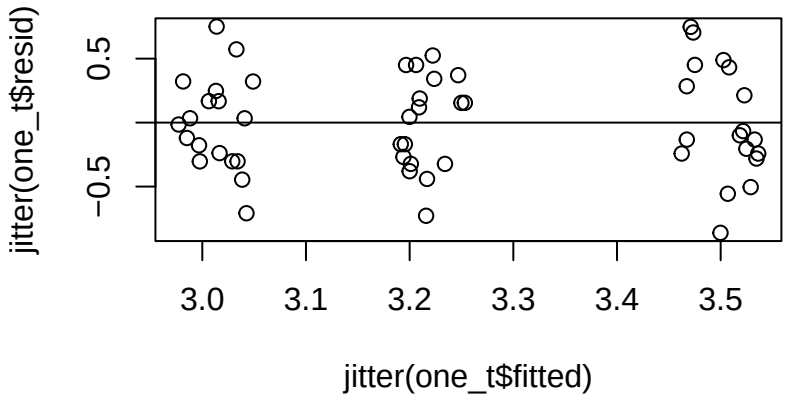
```r
qqnorm(one_t$resid);qqline(one_t$resid)
```



```r
anova(one_t)
```

```
## Analysis of Variance Table
##
## Response: log(breaks)
##           Df Sum Sq Mean Sq F value   Pr(>F)
## tension    2 2.1762 1.08808  6.9894 0.002077 **
## Residuals 51 7.9395 0.15568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(jitter(one_t$fitted),jitter(one_t$resid));abline(h=0)
```



How do the OLS assumptions appear in the transformed model? -

How do the pairwise comparisons for each model compare? Any differences?

3. Two-way ANOVA of `breaks` with both `tension` and `wool` without an interaction

```r
summary(aov(breaks~tension+wool,warpbreaks))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## tension      2   2034  1017.1   7.537 0.00138 **
```

```
## wool         1    451   450.7   3.339 0.07361 .
## Residuals   50   6748   135.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

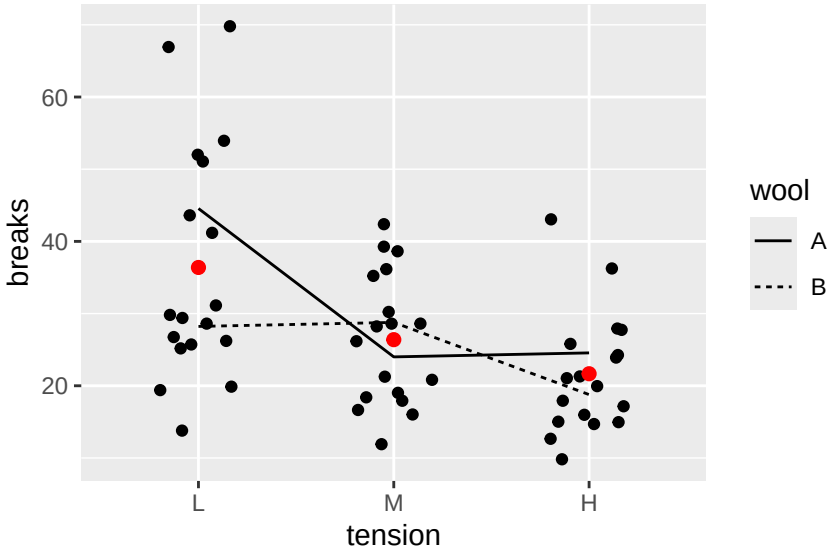
```
summary(lm(breaks~tension+wool,warpbreaks))
```

```
##
## Call:
## lm(formula = breaks ~ tension + wool, data = warpbreaks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.500  -8.083  -2.139   6.472  30.722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.278      3.162  12.423  < 2e-16 ***
## tensionM     -10.000      3.872  -2.582 0.012787 *
## tensionH     -14.722      3.872  -3.802 0.000391 ***
## woolB         -5.778      3.162  -1.827 0.073614 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.62 on 50 degrees of freedom
## Multiple R-squared:  0.2691, Adjusted R-squared:  0.2253
## F-statistic: 6.138 on 3 and 50 DF,  p-value: 0.00123
```

4. Two-way ANOVA of `breaks` with both `tension` and `wool` with an interaction

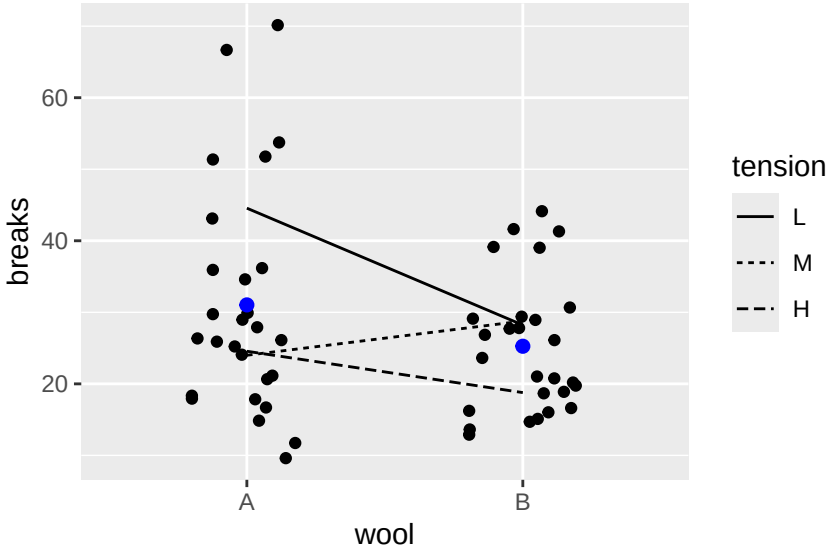
4.A. Graphical understanding

```
ggplot(warpbreaks,aes(x=tension,y=breaks))+
  geom_point(position=position_jitter(width=0.2))+
  stat_summary(fun=mean, geom="point", color="red",size=2)+
  stat_summary(fun=mean, geom="line",
               aes(group=wool, linetype=wool))
```



- When more than 3 categories, it gets hard to read. We still see a decreasing trend from a to b for L and H

```
ggplot(warpbreaks,aes(x=wool,y=breaks))+
  geom_point(position=position_jitter(width=0.2))+
  stat_summary(fun=mean, geom="point", color="blue",size=2)+
  stat_summary(fun=mean, geom="line",
               aes(group=tension, linetype=tension))
```



4.B. Through modeling, add interaction term. : removes main effects, * adds interaction and main effects.

```
aov(breaks~tension*wool,warpbreaks)
```

```
## Call:
##    aov(formula = breaks ~ tension * wool, data = warpbreaks)
##
## Terms:
##                   tension     wool tension:wool Residuals
## Sum of Squares  2034.259  450.667     1002.778  5745.111
## Deg. of Freedom        2        1            2        48
##
## Residual standard error: 10.94028
## Estimated effects may be unbalanced
```

```
aov(breaks~tension+wool+tension*wool,warpbreaks)
```

```
## Call:
##    aov(formula = breaks ~ tension + wool + tension * wool, data = warpbreaks)
##
## Terms:
##                   tension     wool tension:wool Residuals
## Sum of Squares  2034.259  450.667     1002.778  5745.111
## Deg. of Freedom        2        1            2        48
##
## Residual standard error: 10.94028
## Estimated effects may be unbalanced
```

```
aov(breaks~tension+wool+tension:wool,warpbreaks)
```

```
## Call:
##    aov(formula = breaks ~ tension + wool + tension:wool, data = warpbreaks)
##
## Terms:
##                   tension     wool tension:wool Residuals
## Sum of Squares  2034.259  450.667     1002.778  5745.111
## Deg. of Freedom        2        1            2        48
##
## Residual standard error: 10.94028
## Estimated effects may be unbalanced
```

```r
summary(lm(breaks~tension*wool,warpbreaks))
```

```
##
## Call:
## lm(formula = breaks ~ tension * wool, data = warpbreaks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.5556  -6.8889  -0.6667   7.1944  25.4444
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      44.556      3.647  12.218 2.43e-16 ***
## tensionM        -20.556      5.157  -3.986 0.000228 ***
## tensionH        -20.000      5.157  -3.878 0.000320 ***
## woolB           -16.333      5.157  -3.167 0.002677 **
## tensionM:woolB   21.111      7.294   2.895 0.005698 **
## tensionH:woolB   10.556      7.294   1.447 0.154327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.94 on 48 degrees of freedom
## Multiple R-squared:  0.3778, Adjusted R-squared:  0.3129
## F-statistic: 5.828 on 5 and 48 DF,  p-value: 0.0002772
```

```r
aov(breaks~tension:wool,warpbreaks)
```

```
## Call:
##    aov(formula = breaks ~ tension:wool, data = warpbreaks)
##
## Terms:
##                 tension:wool Residuals
## Sum of Squares      3487.704  5745.111
## Deg. of Freedom            5        48
##
## Residual standard error: 10.94028
## 1 out of 7 effects not estimable
## Estimated effects may be unbalanced
```

```r
summary(lm(breaks~tension:wool,warpbreaks))
```

```
##
```

```
## Call:
## lm(formula = breaks ~ tension:wool, data = warpbreaks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.5556  -6.8889  -0.6667   7.1944  25.4444
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      18.778      3.647   5.149 4.84e-06 ***
## tensionL:woolA   25.778      5.157   4.998 8.11e-06 ***
## tensionM:woolA    5.222      5.157   1.013   0.3163
## tensionH:woolA    5.778      5.157   1.120   0.2682
## tensionL:woolB    9.444      5.157   1.831   0.0733 .
## tensionM:woolB   10.000      5.157   1.939   0.0584 .
## tensionH:woolB       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.94 on 48 degrees of freedom
## Multiple R-squared:  0.3778, Adjusted R-squared:  0.3129
## F-statistic: 5.828 on 5 and 48 DF,  p-value: 0.0002772
```

What is the difference between tension*wool and tension:wool?

4.C. Calculate the mean number of breaks by wool and tension.

```
warpbreaks |>  group_by(wool, tension)  |>  summarise(m = mean(breaks))
```

```
## # A tibble: 6 x 3
## # Groups:   wool [2]
##   wool  tension     m
##   <fct> <fct>   <dbl>
## 1 A     L        44.6
## 2 A     M        24
## 3 A     H        24.6
## 4 B     L        28.2
## 5 B     M        28.8
## 6 B     H        18.8
```

Which combination has the lowest mean(breaks)?

Does this give you enough information to say whether the combination with the lowest number of breaks is significantly better than the others?

4.D. Pairwise comparison through Tukey's honest significance difference test

```
# To assess whether one combination is significantly better than another we can conduc
TukeyHSD(aov(breaks~tension*wool,warpbreaks))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = breaks ~ tension * wool, data = warpbreaks)
##
## $tension
##           diff       lwr       upr     p adj
## M-L -10.000000 -18.81965 -1.180353 0.0228554
## H-L -14.722222 -23.54187 -5.902575 0.0005595
## H-M  -4.722222 -13.54187  4.097425 0.4049442
##
## $wool
##          diff       lwr       upr    p adj
## B-A -5.777778 -11.76458 0.2090243 0.058213
##
## $`tension:wool`
##                 diff       lwr        upr     p adj
## M:A-L:A -20.5555556 -35.86188  -5.249234 0.0029580
## H:A-L:A -20.0000000 -35.30632  -4.693678 0.0040955
## L:B-L:A -16.3333333 -31.63966  -1.027012 0.0302143
## M:B-L:A -15.7777778 -31.08410  -0.471456 0.0398172
## H:B-L:A -25.7777778 -41.08410 -10.471456 0.0001136
## H:A-M:A   0.5555556 -14.75077  15.861877 0.9999978
## L:B-M:A   4.2222222 -11.08410  19.528544 0.9626541
## M:B-M:A   4.7777778 -10.52854  20.084100 0.9377205
## H:B-M:A  -5.2222222 -20.52854  10.084100 0.9114780
## L:B-H:A   3.6666667 -11.63966  18.972988 0.9797123
## M:B-H:A   4.2222222 -11.08410  19.528544 0.9626541
## H:B-H:A  -5.7777778 -21.08410   9.528544 0.8705572
## M:B-L:B   0.5555556 -14.75077  15.861877 0.9999978
## H:B-L:B  -9.4444444 -24.75077   5.861877 0.4560950
## H:B-M:B -10.0000000 -25.30632   5.306322 0.3918767
```

```
# the notation of the pairwise comparisons or contrasts can be confusing, you can read
```

4.E. One-way ANOVA of `breaks` with `tension` as a random factor

- lmer pacakge adds a random effect for tension, breaks vary by tension. we are interested in the variance of random effect, we are testing if the variance in the grouping contribute

to the model. In this case, the CI does not contain 0, the variance from tension does contribute to the variability in the outcome.

```
library(lme4)
rmod_fixed <- lm(breaks ~ tension, warpbreaks)
summary(rmod_fixed)
```

```
##
## Call:
## lm(formula = breaks ~ tension, data = warpbreaks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.389  -8.139  -2.667   6.333  33.611
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    36.39       2.80  12.995  < 2e-16 ***
## tensionM      -10.00       3.96  -2.525 0.014717 *
## tensionH      -14.72       3.96  -3.718 0.000501 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.88 on 51 degrees of freedom
## Multiple R-squared:  0.2203, Adjusted R-squared:  0.1898
## F-statistic: 7.206 on 2 and 51 DF,  p-value: 0.001753
```

```
rmod<-lmer(breaks~(1|tension), warpbreaks)
# (1|tension) means we are fitting as random effect
summary(rmod)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: breaks ~ (1 | tension)
##    Data: warpbreaks
##
## REML criterion at convergence: 420.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7882 -0.6811 -0.2867  0.4776  2.9253
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
```

```
##  tension  (Intercept)  48.67    6.976
##  Residual             141.15   11.881
## Number of obs: 54, groups:  tension, 3
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    28.15       4.34   6.486
```

```
confint(rmod, oldNames = FALSE) # oldnames = F tells R to use newer, more informative n
```

```
##                              2.5 %   97.5 %
## sd_(Intercept)|tension  1.516417 17.90831
## sigma                   9.900843 14.61984
## (Intercept)            18.254891 38.04140
```