STA 610 LAB 1

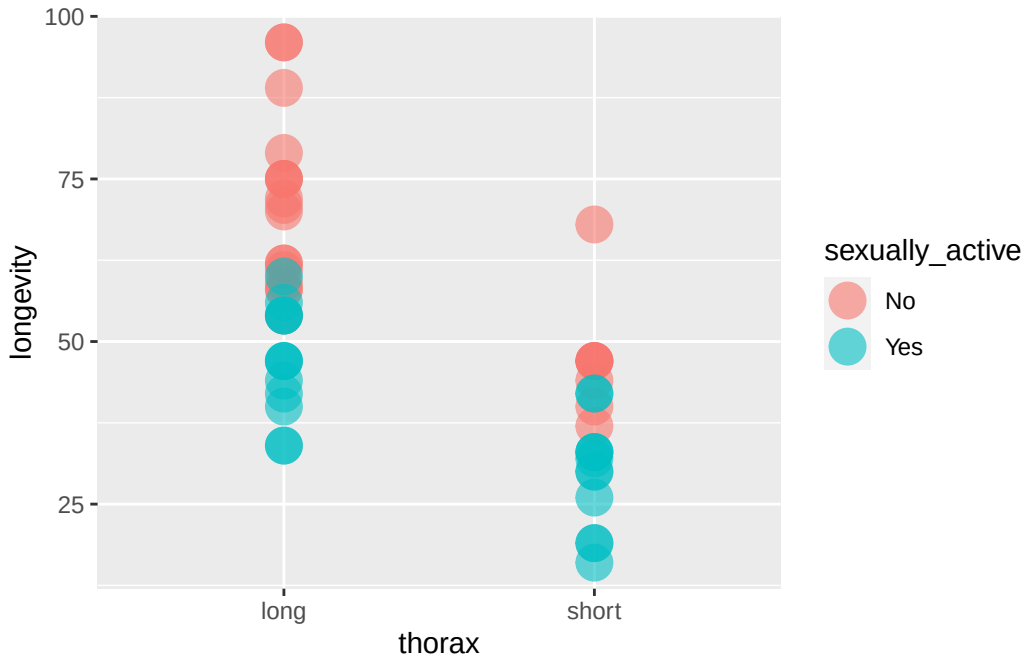
William Tirone

2023-09-06

Ex. 1 : Create a visualization

And interpreting, it looks like sexual activity and a long thorax leads to the longest life, while a shorter thorax leads to a much shorter life regardless sexual activity.

```
ggplot(fly, aes(x=thorax, y=longevity, color=sexually_active)) +
    geom_point(size=6, alpha=0.6)
```



Ex 2. : Two-way ANOVA

Since we have the variables j, `sexually_active = ('No', 'Yes',)` for j = 1 or 2 and k `thorax = ('long','short')` for k = 1 or 2, we can consider:

$$y_{ijk} = \mu + \alpha I(j = 2) + \beta I(k = 2) + \gamma I(j = k = 2) + \epsilon_{ijk}$$

```
aov.fly = aov(longevity ~ sexually_active * thorax, data = fly)
coef(summary.lm(aov.fly))
```

```
                                Estimate Std. Error   t value     Pr(>|t|)
(Intercept)                    71.294118   2.504986 28.460888 7.515494e-31
sexually_activeYes            -24.140271   3.805346 -6.343778 8.823222e-08
thoraxshort                   -24.169118   4.428231 -5.457962 1.863736e-06
sexually_activeYes:thoraxshort  6.598605   6.058420  1.089163 2.817554e-01
```

So now we can finish our model and say that:

$$y_{ijk} = 71.29 - 24.17 I(j = 2) - 24.14 I(k = 2) + 6.6 I(j = k = 2) + \epsilon_{ijk}$$

And considering whether or not we have statistical significance for the coefficients:

```
summary(aov.fly)
```

```
                     Df Sum Sq Mean Sq F value   Pr(>F)
sexually_active       1   7713    7713  72.303 5.42e-11 ***
thorax                1   4978    4978  46.662 1.63e-08 ***
sexually_active:thorax  1    127     127   1.186    0.282
Residuals            46   4907     107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ex. 3

To estimate fruitfly lifespan in days (longevity), we have considered two binary variables: sexual activity (yes or no), and thorax length (long or short). We also include a so-called "interaction effect" which indicates that sexual activity and thorax length interact with each other in a way not explained by either variable alone.

Looking at our table above from example 2, we see that both sexual activity and thorax reach statistical significance on their own, but with a large p-value of 0.282, the interactive between the two does not have statistical significance. However, we also have to quantify uncertainty which we do using the output from our ANOVA model. These confidence interval can be viewed here:

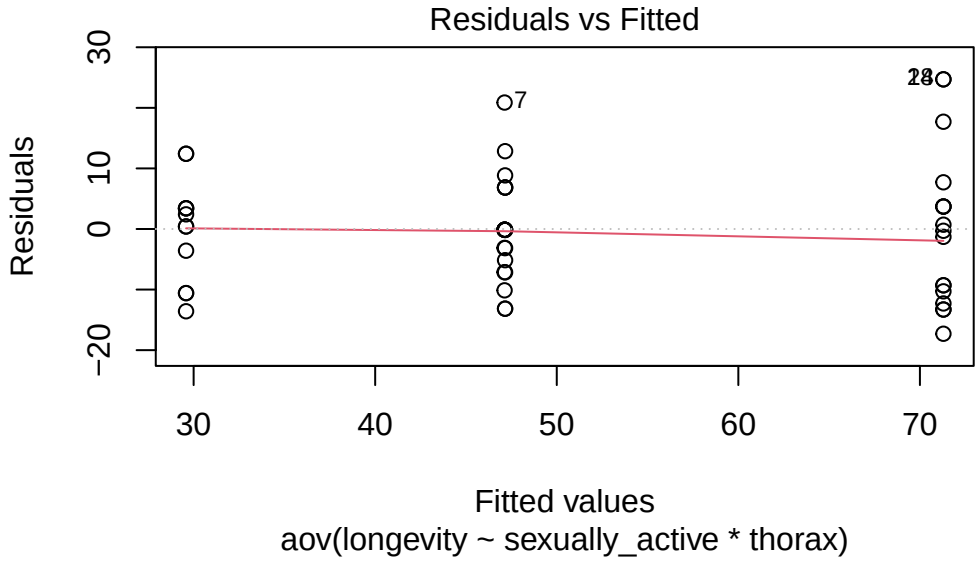
```
confint(aov.fly)
```

```
                                  2.5 %    97.5 %
(Intercept)                    66.251843  76.33639
sexually_activeYes            -31.800037 -16.48051
thoraxshort                   -33.082684 -15.25555
sexually_activeYes:thoraxshort  -5.596362  18.79357
```

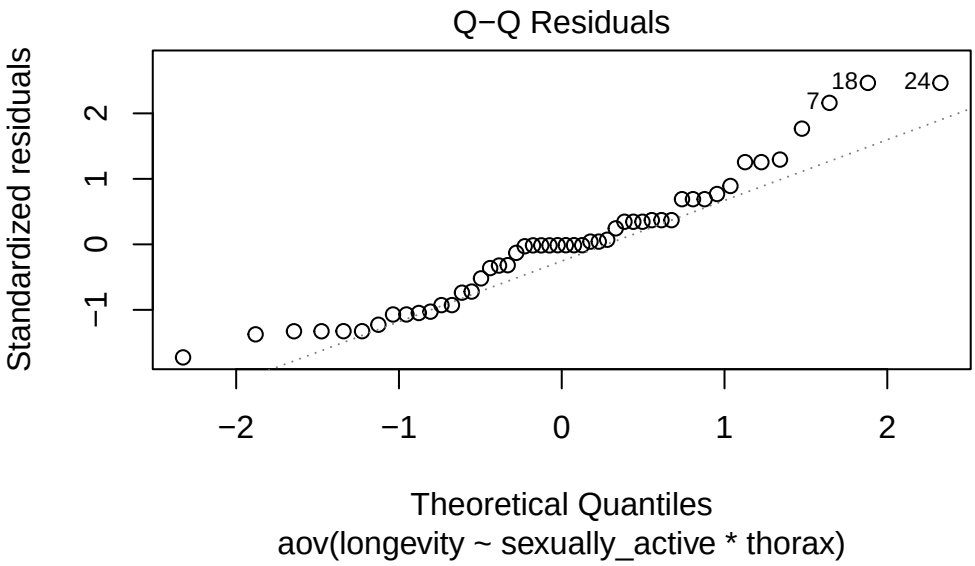
Ex. 4

Considering two diagnostic plots below, we don't necessarily see any strong patterns in the residuals. However, the Q-Q plot shows that the residuals look very non-normal, so it is likely our assumption about normal data is broken. Our model is probably not very good since we are violating many assumptions!

```
plot(aov.fly, 1)
```



```r
plot(aov.fly, 2)
```



Considering a couple extra tests below, we do not reject the null of the Levene test that all the variances are equal, so we have probably not violated the homoscedasticity assumption.

And with the Shapiro-Wilk test, we can check the null that the residuals are normal. Since this is rejected with a small p-value, we are reassured that the residuals are not normal.

```r
# check variances
leveneTest(longevity ~ sexually_active * thorax, data = fly)
```

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  1.5688 0.2097
      46
```

```r
# check residuals
aov_residuals = residuals(aov.fly)
shapiro.test(x = aov_residuals )
```

```
    Shapiro-Wilk normality test
```

```
data:  aov_residuals
W = 0.95177, p-value = 0.04033
```