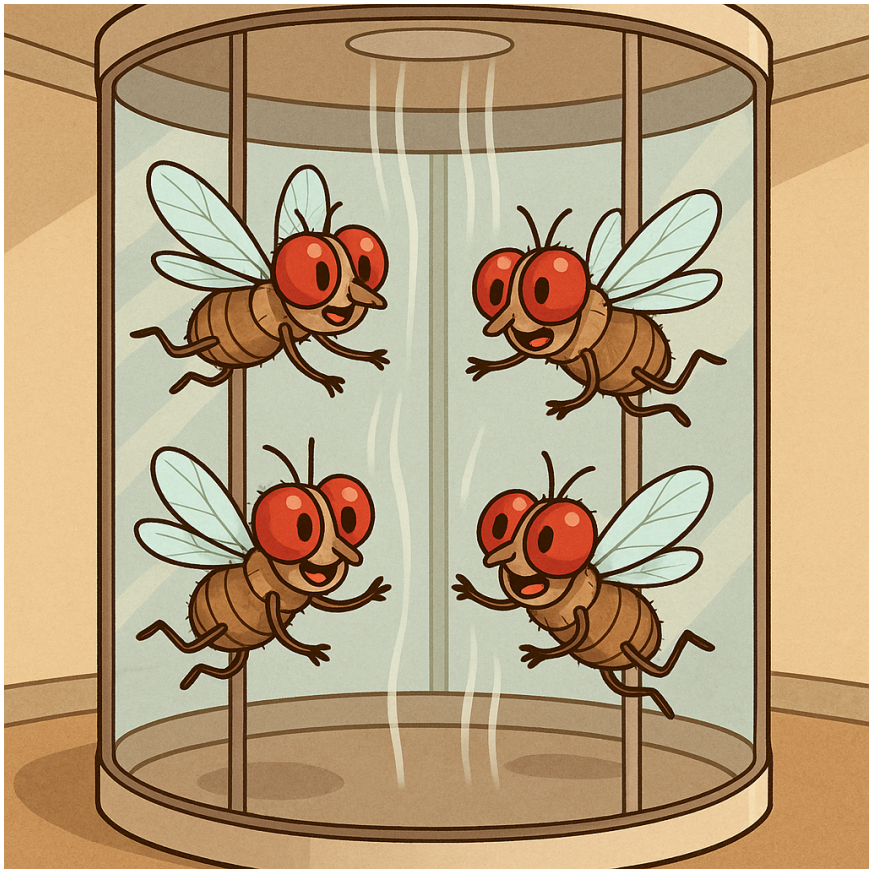
Statistical Analysis of Flight Assay Data in R

Jun Ishigohoka



Environment

We recommend you use `Rstudio`.

- Download `flight_assay_analysis.Rmd` from https://github.com/PallaresLab/EMBL_Drosophila_course_2025/blob/main/scripts/flight_assay_analysis.Rmd, place it in the same directory as your flight assay data.
- Open `flight_assay_analysis.Rmd` in `Rstudio`.
- Follow the instructions

Or

- PDF: download from here
- HTML: access here

Preparation

If you don't have the packages, install them.

```
packages <- c("ggplot2", "car")
for (pkg in packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)
  }
}
```

Load libraries.

```
library(ggplot2)
library(car) # Anova() and leveneTest()
```

Read data in two data frames. For this protocol, test data will be used.

```
d_1 <- read.csv("data/test_data/R4_female.csv")
d_2 <- read.csv("data/test_data/R4_male.csv")

head(d_1)
```

```
##   X.1   Area       X      Y
## 1   1  5.456 407.768  9.192
## 2   2  6.666 343.943 10.074
## 3   3 11.087 306.180 11.690
## 4   4  4.932 376.921 16.866
## 5   5  4.569 148.680 25.018
## 6   6  5.389 330.855 26.006
```

Add a column for the factor you are comparing.

```r
d_1$sex <- "female"
d_2$sex <- "male"
```

Concatenate the two data frames. Then set it as a factor.

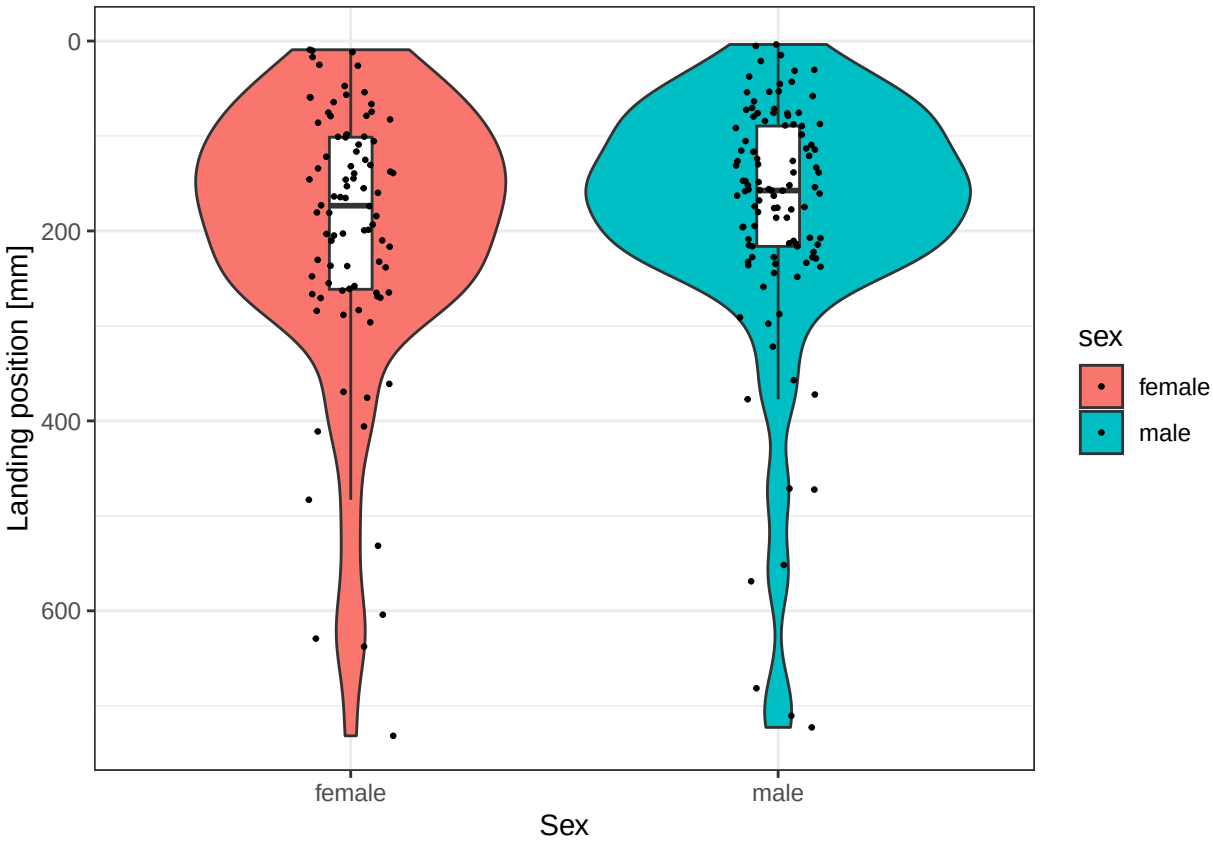
```r
d <- rbind(d_1, d_2)
d$sex <- factor(d$sex, levels = c("female", "male"))
```

Visualisation

Let's visualise the distribution of landing height. **Once you have visualised the distributions, try to describe them.**

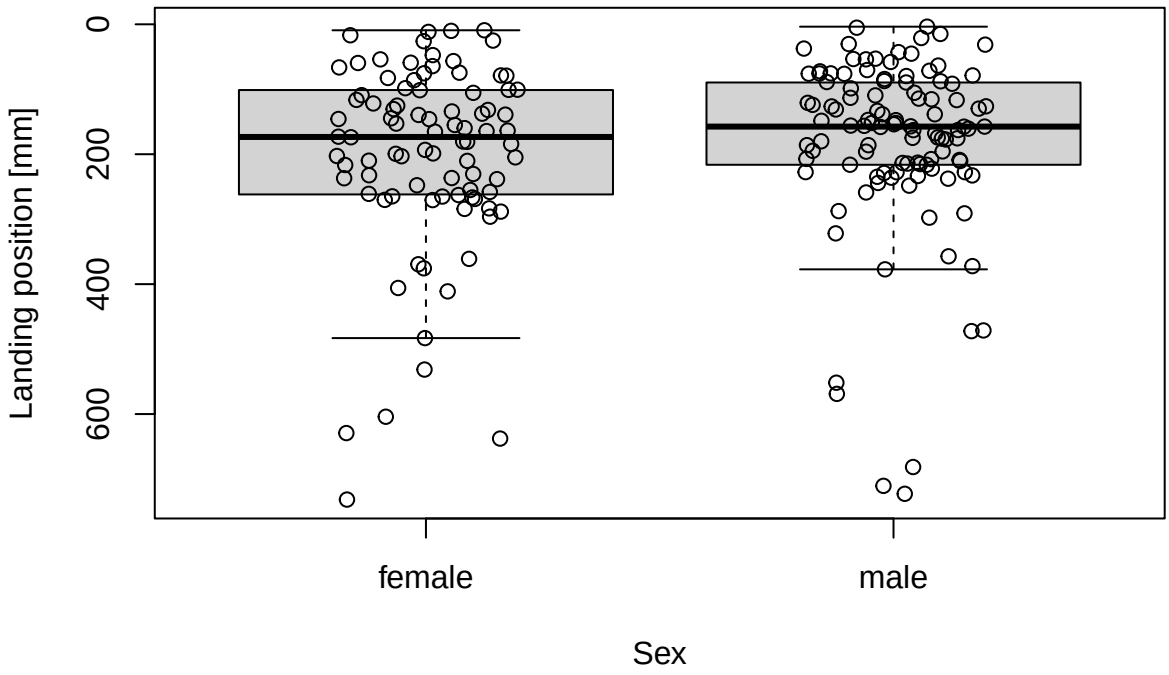
If you have `ggplot2`...

```r
ggplot(data = d, # data to be plotted
       mapping = aes(x=sex, y=Y, fill = sex)) + # columns in the data frame to be used
        geom_violin() +   # violin plot
        geom_boxplot(     # box plot
          width = 0.1,    # width of the box plot
          fill = "white", # colour inside the box plot
          outliers = F    # whether to have points for outliers.
                          # FALSE because of geom_fitter() below
          ) +
        geom_jitter(      # each data point with "jitter" along x axis
          size = 0.5,     # size of points
          width = 0.1     # the amount of jitter
          ) +
        scale_y_reverse(  # to make the plot same direction as the tapes
          limits = c(NA, 0) # so that the lowest Y is 0
          ) +
  labs(                   # Add labels
    x = "Sex",
    y = "Landing position [mm]") +
  theme_bw()
```



If you prefer the base plot or do not have `ggplot2` installed.

```
boxplot(d$Y ~ as.factor(d$sex),
     pch = NA,
     xlab = "Sex",
     ylab = "Landing position [mm]",
     ylim = rev(range(d$Y))
)
points(jitter(as.numeric(as.factor(d$sex))), d$Y)
```



Statistical analysis (hypothesis testing)

Are your data normally distributed?

Many parametric tests assume normality. Let's run Shapiro-Wilk test for each group to test normality.

```
shapiro.test(subset(d, sex == "female")$Y)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(d, sex == "female")$Y
## W = 0.87191, p-value = 3.497e-07
```

```
shapiro.test(subset(d, sex == "male")$Y)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset(d, sex == "male")$Y
## W = 0.80816, p-value = 1.326e-10
```

Does the result fit your description of the distributions?

If they are not normally distributed, think why. Do you expect normally distributed data from this experiment? (Hints: Poisson process; waiting time)

Does your data have equal variance?

Many parametric tests assume equal variance between groups.

Let's run two tests for homogeneity: F-test and Brown-Forsythe test.

Run F-test.

```
var.test(Y ~ sex, d)
```

```
##
##  F test to compare two variances
##
## data:  Y by sex
## F = 1.1212, num df = 87, denom df = 108, p-value = 0.5699
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7535657 1.6828055
## sample estimates:
## ratio of variances
##           1.121162
```

Run Brown-Forsythe test (`car` needs to be installed).

```
leveneTest(Y ~ sex, d, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.8082 0.3698
##       195
```

If you do not have `car` installed, you can run Brown-Forsythe test manually. Compute absolute deviance from median for each sex manually, and run ANOVA.

```
d$z <- NA
for (sex in levels(d$sex)) {
  idx <- which(d$sex == sex)
  d[idx,]$z <- abs(d[idx, ]$Y - median(d[idx, ]$Y))
}

summary(aov(d$z ~ d$sex))
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## d$sex         1    9114    9114   0.808   0.37
## Residuals   195 2199050   11277
```

Is the mean landing distance different between groups?

If the data are normally distributed and have equal variance, student t-test

```r
t.test(d$Y ~ d$sex,
       var.equal = T,
       paired = F
       )
```

```
##
##  Two Sample t-test
##
## data:  d$Y by d$sex
## t = 0.92975, df = 195, p-value = 0.3537
## alternative hypothesis: true difference in means between group female and group male
## 95 percent confidence interval:
##  -20.94187  58.29702
## sample estimates:
## mean in group female   mean in group male
##             200.7005             182.0229
```

If data are normally distributed but not with equal variance, Welch's t-test

```r
t.test(d$Y ~ d$sex,
       var.equal = F,
       paired = F
       )
```

```
##
##  Welch Two Sample t-test
##
## data:  d$Y by d$sex
## t = 0.92407, df = 181.57, p-value = 0.3567
## alternative hypothesis: true difference in means between group female and group male
## 95 percent confidence interval:
##  -21.20353  58.55868
## sample estimates:
## mean in group female   mean in group male
##             200.7005             182.0229
```

If data are not normally distributed, you can run non-parametric tests or permutation test.

Option 1: Non-parametric test

Let's run Mann–Whitney U test.

```r
wilcox.test(d$Y ~ d$sex,
            paired = F
            )
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  d$Y by d$sex
## W = 5247, p-value = 0.2575
## alternative hypothesis: true location shift is not equal to 0
```

Option 2: Permutation test

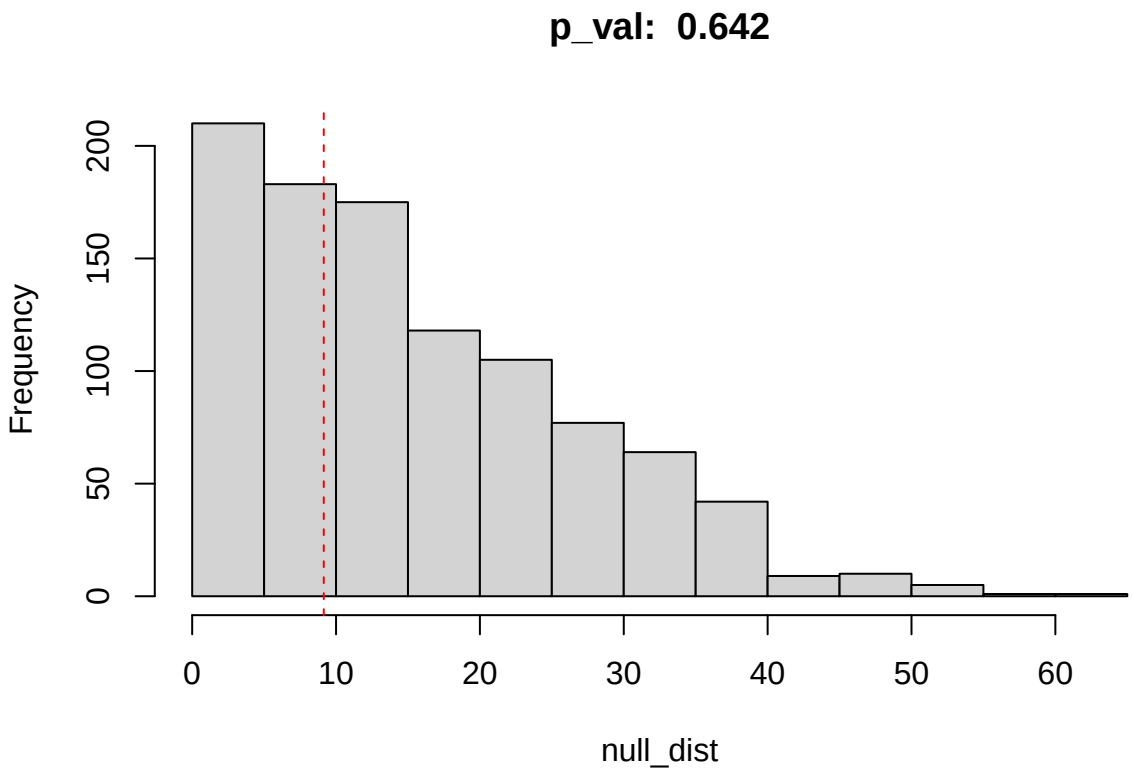
Let's write a function to run a permutation test. Briefly, we shuffle the label, compute difference between groups. We repeat this process many times to get a null distribution. We compute p-value based on the position of the observed difference between groups in the null distribution.

```r
# Null distribution of abs(mean(Y | female) - mean(Y | male))
null_dist <- sapply(1:1000, # repeat 1000 times
       function(x){
         d_tmp <- d
         d_tmp$sex <- sample(d$sex) # shuffle label
         return(abs(mean(d_tmp[d_tmp$sex == "female","Y"])
                  - mean(d_tmp[d_tmp$sex == "male","Y"]))
           )
       }
       )

# Observed value of mean(female) - mean(male)
obs <- abs(mean(d[d$sex == "female", "Y"] - d[d$sex == "male", "Y"]))

# p value as the rank of observation in the null distribution
p_val <- sum(null_dist > obs )/ length(null_dist)

hist(null_dist,
     main = paste("p_val: ", p_val))
abline(v = obs, col = "red", lty = 2)
```



Optional: Statistical modeling

Here, we conduct statistical modeling of landing distance using a generalised linear model (GLM) approach with the exponential distribution, considering landing events as a Poisson process with a constant "landing rate". Because the waiting time of a Poisson process is an exponentially distributed random variable, we can use a GLM with exponential distribution. We compare the result with a linear model (LM), which assumes a normal distribution as the underlying distribution.

GLM

Fit the data to an exponential GLM.

```
glm_1 <- glm(Y ~ sex,
             data = d,
             family = Gamma(link = "log") # This is how to specify exponential
             )
```

According to the fitted GLM, do the two groups have different landing rate?

If you have `car` installed

```
Anova(glm_1)
```

```
## Analysis of Deviance Table (Type II tests)
```

```
##
## Response: Y
##     LR Chisq Df Pr(>Chisq)
## sex  0.85803  1     0.3543
```

```
summary(glm_1, dispersion = 1) # exponential is a Gamma with dispersion = 1
```

```
##
## Call:
## glm(formula = Y ~ sex, family = Gamma(link = "log"), data = d)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.30181    0.10660  49.735   <2e-16 ***
## sexmale     -0.09768    0.14331  -0.682    0.495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 1)
##
##     Null deviance: 111.53  on 196  degrees of freedom
## Residual deviance: 111.07  on 195  degrees of freedom
## AIC: 2427.1
##
## Number of Fisher Scoring iterations: 5
```

If you do not have `car` installed

```
drop1(glm_1, test = "Chi")
```

```
## Single term deletions
##
## Model:
## Y ~ sex
##        Df Deviance    AIC scaled dev. Pr(>Chi)
## <none>      111.07 2427.1
## sex     1   111.53 2425.9     0.85803   0.3543
```

LM

Fit the data to a linear model.

```
lm_1 <- lm(Y ~ sex,
           data = d)
```

If you do not have `car` installed

```r
drop1(lm_1, test = "Chi")
```

```
## Single term deletions
##
## Model:
## Y ~ sex
##        Df Sum of Sq     RSS    AIC Pr(>Chi)
## <none>              3831693 1949.5
## sex     1     16986 3848679 1948.4   0.3506
```

```r
summary(lm_1)
```

```
##
## Call:
## lm(formula = Y ~ sex, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -191.51  -94.30  -25.02   45.70  540.72
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   200.70      14.94   13.43   <2e-16 ***
## sexmale       -18.68      20.09   -0.93    0.354
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 140.2 on 195 degrees of freedom
## Multiple R-squared:  0.004413,   Adjusted R-squared:  -0.0006922
## F-statistic: 0.8644 on 1 and 195 DF,  p-value: 0.3537
```

GLM vs LM

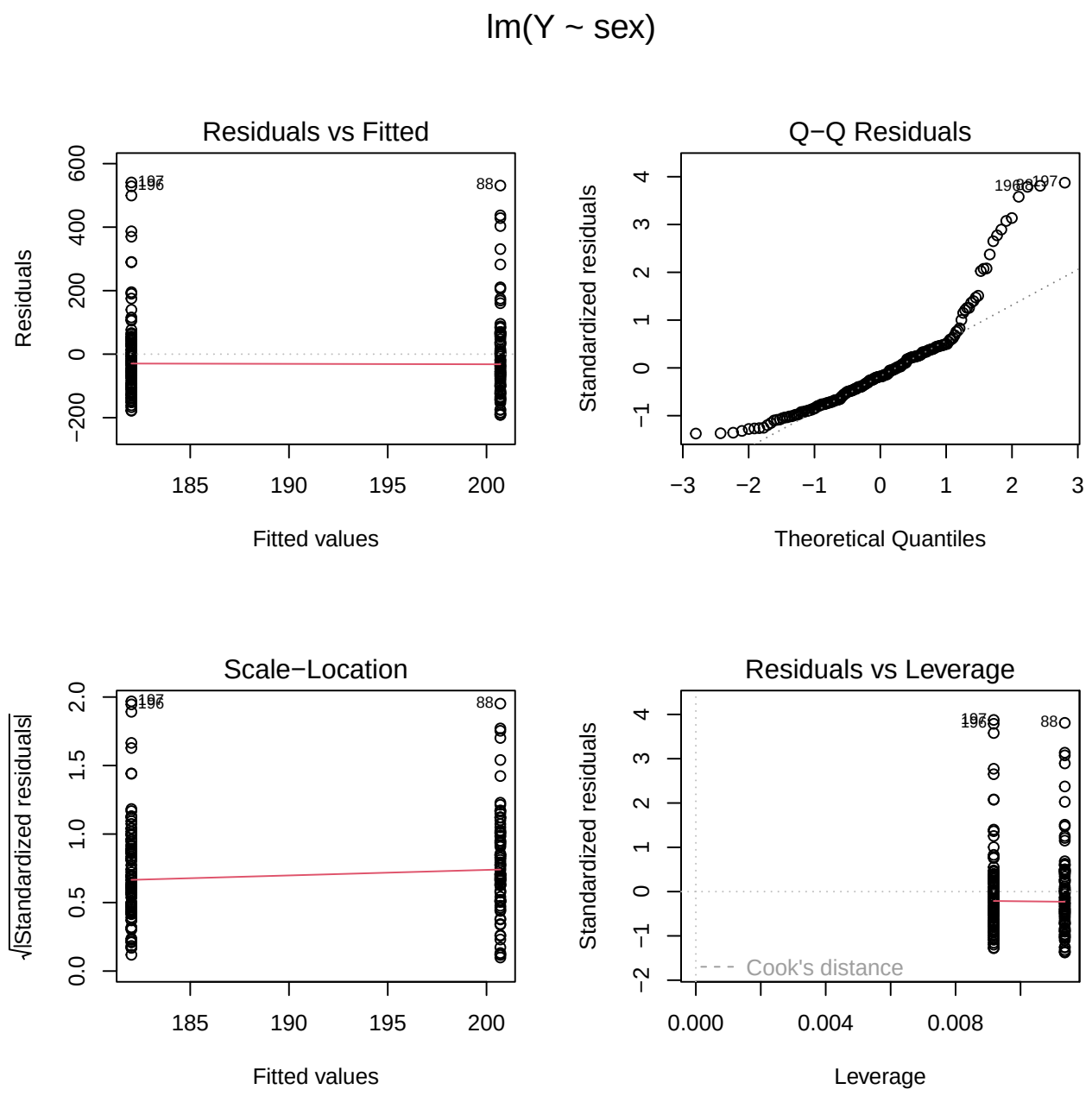
Compare the diagnostic plots[1] between `glm_1` and `lm_1`. **Is using LM for landing distance data worse than exponential GLM?**

- Is the linearity assumption met?
- Is the data homoscedastic?
- Are residuals distributed normally?
- Are there outliers?

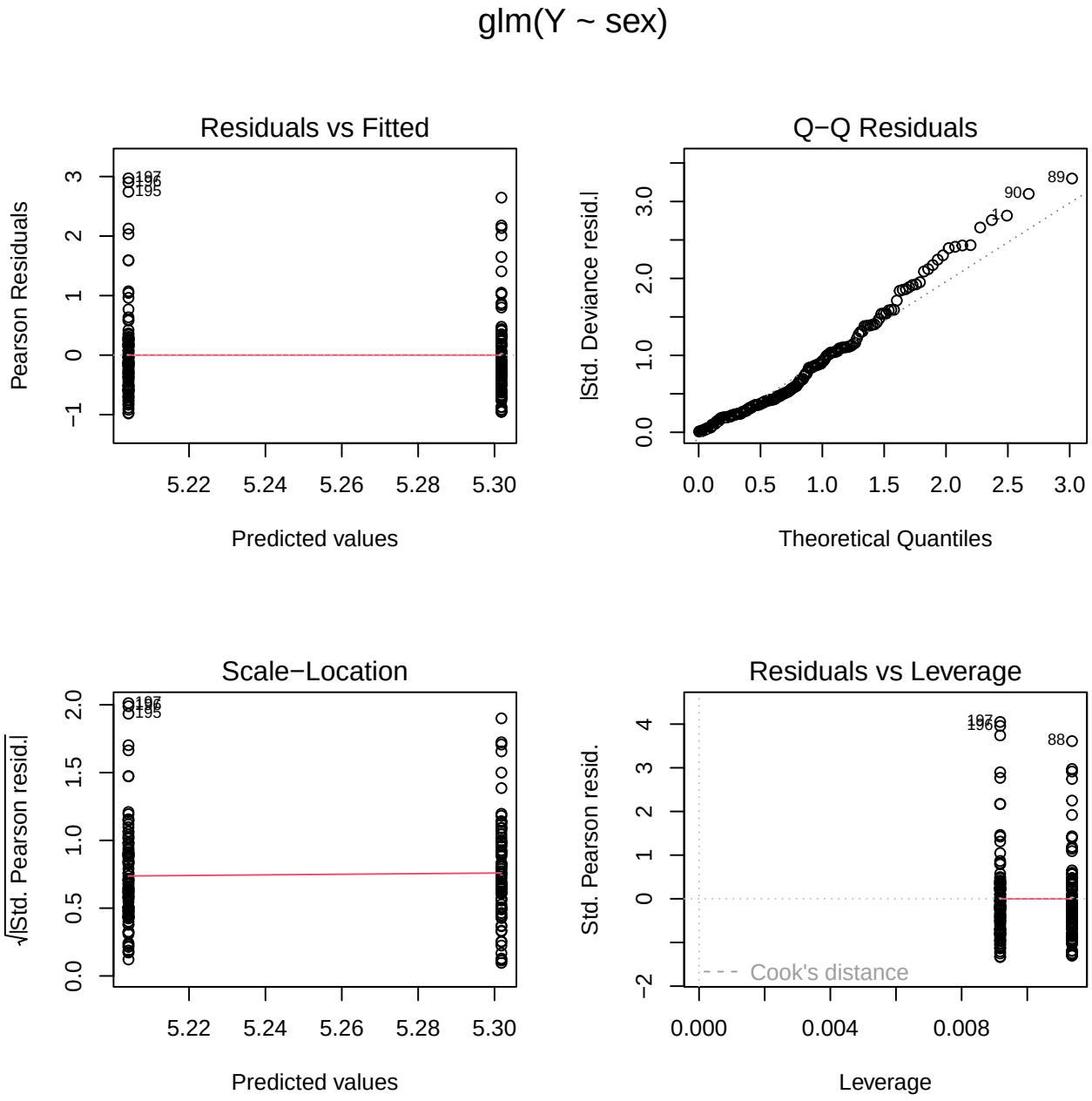
```r
par(mfrow = c(2,2),
    oma = c(0, 0, 2, 0)
```

[1] What do the diagnostic plots mean? https://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/

```
    )
plot(lm_1)
```




```
par(mfrow = c(2,2),
    oma = c(0, 0, 2, 0)
    )
plot(glm_1)
```



What's next?

In real life, we have multiple biological and technical replicates. **How do we account for such random effects?**

Further reading

- Whitlock & Schluter. *The Analysis of Biological Data 3rd edition.* Macmillan Learning. 2020
- Zuur et al. *Mixed Effects Models and Extensions in Ecology with R.* Springer. 2009.