simon-5502-04-slides

Topics to be covered

- What you will learn
  - Review one factor analysis of variance
  - Multiple factor analysis of variance
  - Checking assumptions of analysis of variance
  - Interactions in analysis of variance

Review oneway analysis of variance

- $H_0 : \ \mu_1 = \mu_2 = \ldots = \mu_k$
- $H_1 : \ \mu_i \neq \mu_j$ for some i, j
  - Reject H_0 if F-ratio is large
- Note: when k=2, use analysis of variance or t-test

Speaker notes

In Biostats-1, we discussed the comparison of three or more means using oneway or single factor analysis of variance. You can actually use analysis of variance when comparing only two means, but an equivalent alternative is the t-test.

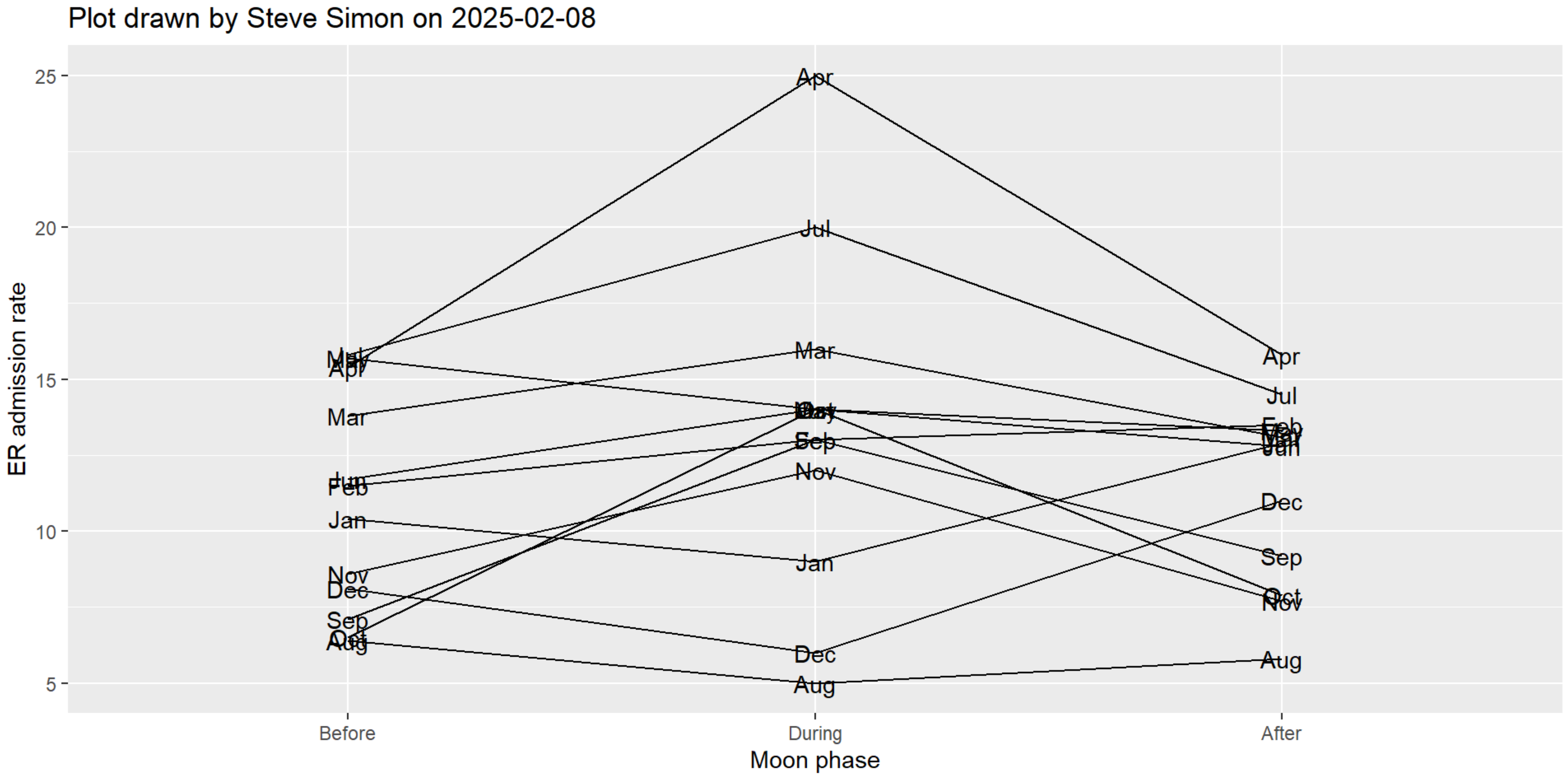
Full moon data

- Admission rates to mental health clinic
  - Before, during, and after full moon.
- One year of data, starting in August
- Consult the data dictionary for more details

Speaker notes

To illustrate oneway analysis of variance, I found a dataset on mental health clinic admissions.

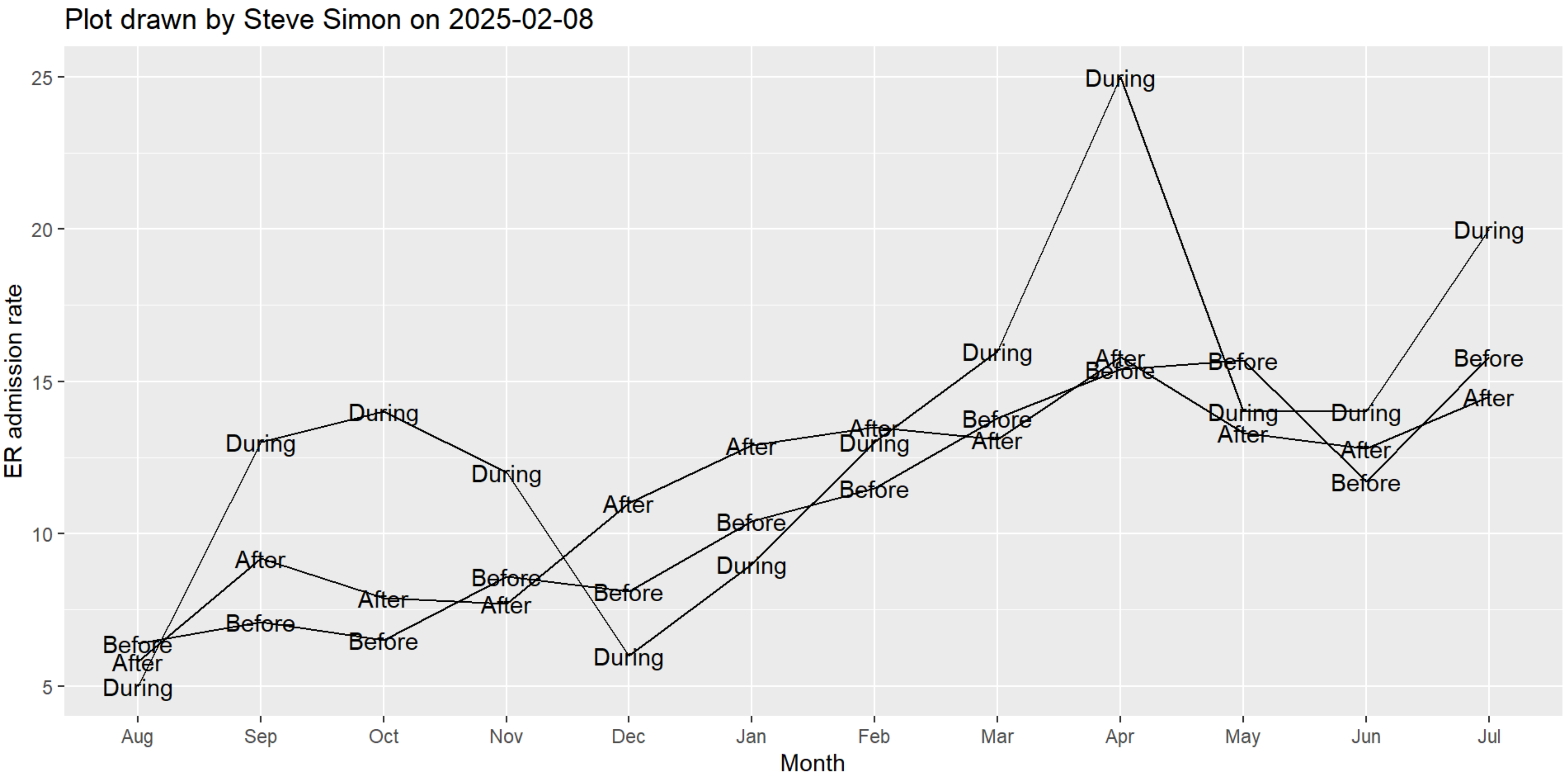
Line graph of full moon data



Speaker notes

For interpretation of this output from R and all other output from R included below, refer to the file simon-5502-04-demo.

Alternative line graph of full moon data



Descriptive statistics

```
# A tibble: 3 × 3
  Moon   Admission_mean Admission_sd
  <fct>           <dbl>        <dbl>
1 Before           10.9         3.62
2 During           13.4         5.50
3 After            11.5         3.11
```

Analysis of variance table

```
# A tibble: 2 × 7
  term             df.residual   rss    df sumsq statistic p.value
  <chr>                  <dbl> <dbl> <dbl> <dbl>     <dbl> <glue>
1 Admission ~ 1             35  625.    NA  NA        NA   <NA>
2 Admission ~ Moon          33  583.     2  41.5       1.17 p = 0.322
```

Parameter estimates

```
# A tibble: 3 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl> <glue>
1 (Intercept)   10.9        1.21     8.99  p < 0.001
2 MoonDuring     2.50       1.72     1.46  p = 0.155
3 MoonAfter      0.542      1.72     0.316 p = 0.754
```

Tukey post hoc

```
# A tibble: 3 × 7
  term  contrast      null.value estimate conf.low conf.high adj.p.value
  <chr> <chr>              <dbl>    <dbl>    <dbl>     <dbl> <glue>
1 Moon  During-Before          0    2.50     -1.71      6.71 p = 0.325
2 Moon  After-Before           0    0.542    -3.67      4.75 p = 0.947
3 Moon  After-During           0   -1.96     -6.17      2.25 p = 0.496
```

Live demo, Review one factor analysis of variance

Live demonstration of part 1 of simon-5502-04-demo.qmd

Break #1

- What you have learned
  - Review one factor analysis of variance
- What's coming next
  - Multiple factor analysis of variance

Mathematical model, 1

- Decompose μ_{ij} into $\mu + \alpha_i + \beta_j$
  - α_i is the deviation for the ith level of first factor
  - β_j is the deviation for the jth level of second factor
  - Require $\alpha_1 = 0$ and $\beta_1 = 0$
  - μ is the mean for the reference levels

Speaker notes

The mathematical model for two factor analysis of variance is a bit more complex than a single factor analysis of variance. You have a mean at the reference levels, mu, You also have deviations from the overall mean associated with the first factor (alpha), deviations from the overall mean associated with the second factor (beta)

There are a total of a and b categories for the two categorical independent variables.

Mathematical model, 2

- $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$
  - i=1,…,a levels of the first categorical variable
  - j=1,…,b levels of the second categorical variable
  - k=1,…,n replicates with first and second categories
- Note: $\mu, \alpha_i, \beta_j, \epsilon_{ijk}$ are population values

Speaker notes

The mathematical model for two factor analysis of variance is a bit more complex than a single factor analysis of variance. You have an overall mean, mu, and deviations from the overall mean associated with the first factor (alpha), deviations from the overall mean associated with the second factor (beta) and an error term (epsilon).

There are a total of a and b categories for the two categorical independent variables.

Mathematical model, 3

- $H_0 : \alpha_i = 0$ for all i

- $H_0 : \beta_j = 0$ for all j

Speaker notes

There are two hypotheses. The first, testing that all the alphas equal zero is effectively testing whether the first factor has no impact on the outcome. Testing that all the betas equal zero is effectively testing whether the second factor has no impact on the outcome.

Parameter estimates for the two factor model

```
# A tibble: 14 × 5
   term        estimate std.error statistic p.value
   <chr>          <dbl>     <dbl>     <dbl> <glue>
 1 (Intercept)    4.72      1.50      3.14  p = 0.005
 2 MoonDuring     2.5       0.984     2.54  p = 0.019
 3 MoonAfter      0.542     0.984     0.550 p = 0.588
 4 MonthSep       4.03      1.97      2.05  p = 0.053
 5 MonthOct       3.73      1.97      1.90  p = 0.071
 6 MonthNov       3.70      1.97      1.88  p = 0.073
 7 MonthDec       2.63      1.97      1.34  p = 0.195
 8 MonthJan       5.03      1.97      2.56  p = 0.018
 9 MonthFeb       6.93      1.97      3.52  p = 0.002
10 MonthMar       8.57      1.97      4.35  p < 0.001
11 MonthApr      13.0       1.97      6.61  p < 0.001
12 MonthMay       8.60      1.97      4.37  p < 0.001
13 MonthJun       7.10      1.97      3.61  p = 0.002
```

Analysis of variance table comparing the two factor model to the null model

```
# A tibble: 2 × 7
  term                      df.residual   rss    df sumsq statistic p.value
  <chr>                           <dbl> <dbl> <dbl> <dbl>     <dbl> <glue>
1 Admission ~ 1                      35  625.    NA   NA       NA    <NA>
2 Admission ~ Moon + Month           22  128.    13  497.       6.58 p < 0.001
```

Speaker notes

Analysis of variance table comparing the two factor model to the one factor model

```
# A tibble: 2 × 7
  term                      df.residual   rss    df sumsq statistic p.value
  <chr>                           <dbl> <dbl> <dbl> <dbl>     <dbl> <glue>
1 Admission ~ Moon                   33  583.    NA   NA        NA    <NA>
2 Admission ~ Moon + Month           22  128.    11  456.      7.13 p < 0.001
```

R-squared values

```
# A tibble: 3 × 3
  model  r.squared deviance
  <glue>     <dbl>    <dbl>
1 m1        0          625.
2 m2        0.0664     583.
3 m3        0.795      128.
```

Tukey post hoc test

```
# A tibble: 3 × 7
  term  contrast      null.value estimate conf.low conf.high adj.p.value
  <chr> <chr>              <dbl>    <dbl>    <dbl>     <dbl> <chr>      
1 Moon  After-Before           0    0.542  -1.93        3.01 0.847      
2 Moon  During-Before          0    2.50    0.0280      4.97 0.047      
3 Moon  During-After           0    1.96   -0.514       4.43 0.138      
```

Speaker notes

Use the Tukey posthoc test because the sample sizes are equal across the moon phases. The results are a bit ambiguous because before and after are not statistically different, after and during are not statistically different but before and during are statistically different. This is probably due to a lack of precision and an extra year’s worth of data would help quite a bit.

The analogy I use is travel time. My wife and I live in Leawood. Our son lives in Lee’s Summit. A repair shop we all use is in Olathe. It is not far from Leawood to Olathe. It is not far from Leawood to Lee’s Summit. But it is far from Lee’s Summit to Olathe.

Live demo, Multiple factor analysis of variance

Live demonstration of part 2 of simon-5502-04-demo.qmd

Break #2

- What you have learned
  - Multiple factor analysis of variance
- What’s coming next
  - Checking assumptions of analysis of variance

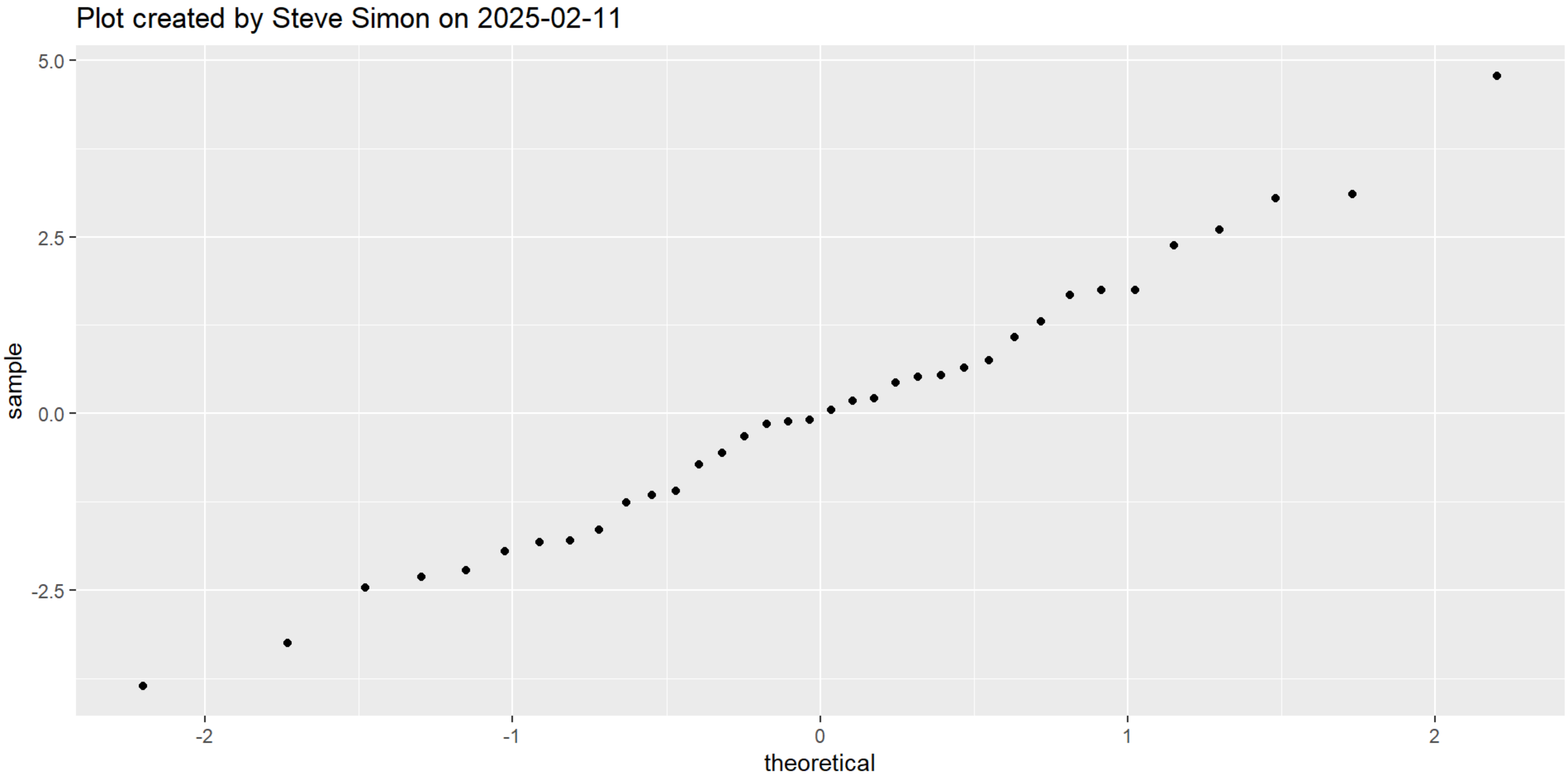
Assumptions

- Normality
- Equal variances
- Independence
- Note: No linearity assumption
  - Only for linear regression and analysis of covariance

Speaker notes

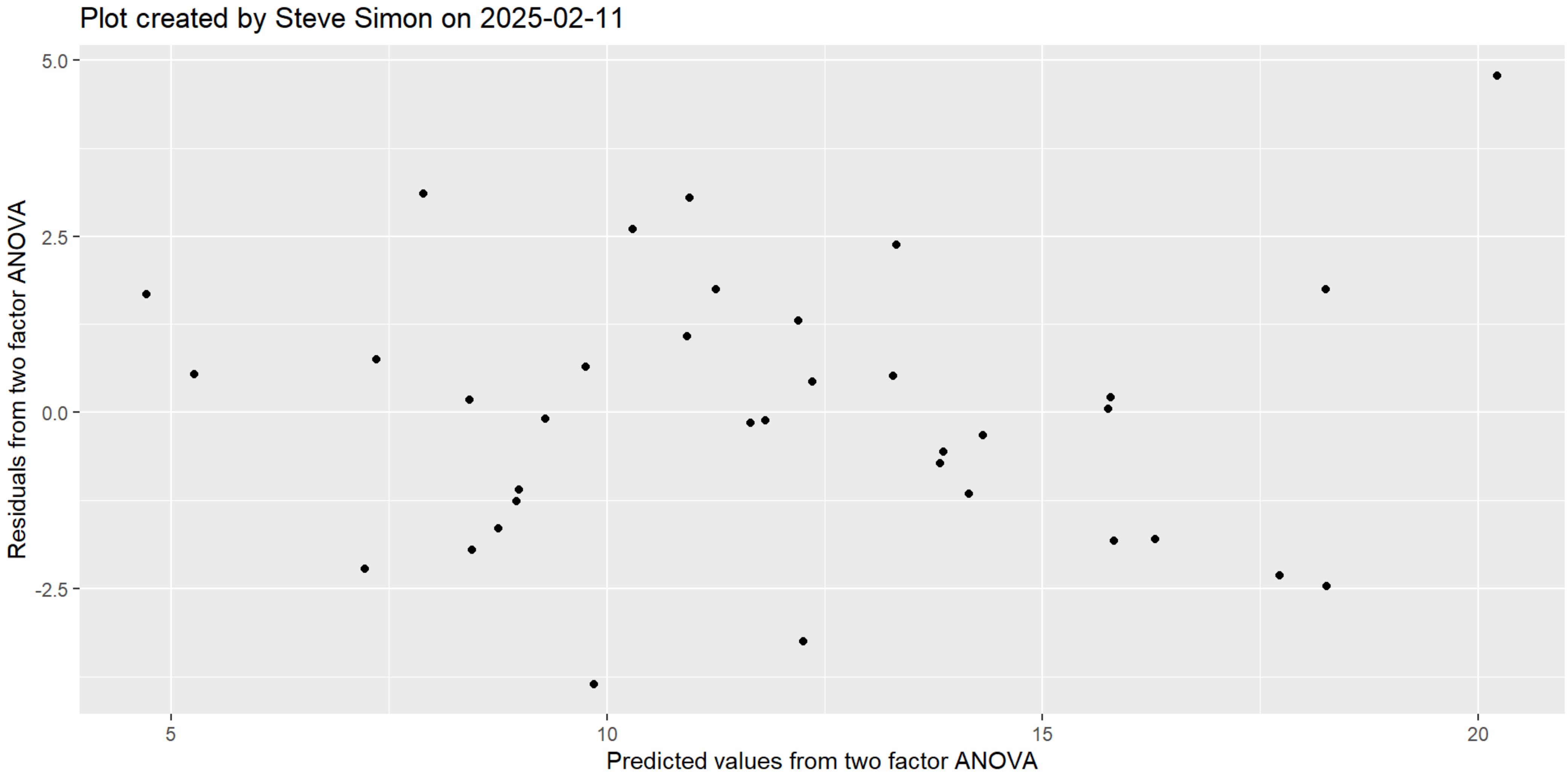
The assumptions for multiple factor analysis of variance are no different than single factor analysis of variance. You must use residuals to check the assumptions of normality and equal variances. The assumption of independence is usually assessed qualitatively.

Q-Q plot of residuals



Plot created by Steve Simon on 2025-02-11

Residual versus predicted value plot



Diagnostic measures that are not needed

- Variance inflation factor
- Leverage
- Studentized deleted residuals
- Cook’s distance

Speaker notes

There are some diagnostic measures that are important in multiple linear regression that are not at all needed for multiple factor analysis of variance. When your independent variables are categorical, there is little opportunity for collinearity to appear. So the variance inflation factor is not normally computed for multiple factor analysis of variance.

While studentized deleted residuals could be used, they almost never deviate substantially from the plain resisuals, so there is no need to look at them.

It is also difficult to produce outliers among the independent variables because categorical data can’t have extreme values. In theory, it could happen if the distribution of data within certain categories was extremely unbalanced. The amount of imbalance, though, before this becomes an issue is never seen in real world datasets.

for the same reason, Cook’s distance, which is a measure that combines leverage with studentized deleted residuals, is never used in multiple factor analysis of variance.

Live demo, Checking assumptions of analysis of variance

Live demonstration of part 3 of simon-5502-04-demo.qmd

Break #3

- What you have learned
    - Checking assumptions of analysis of variance
- What’s coming next
    - Interactions in analysis of variance

What is an interaction

- Impact of one variable is influenced by a second variable
- Example, influence of alcohol on sleeping pills
- Three types of interactions
  - Between two categorical predictors
  - Between a categorical and a continuous predictor
  - Between two continuous predictors
- Interactions greatly complicate interpretation

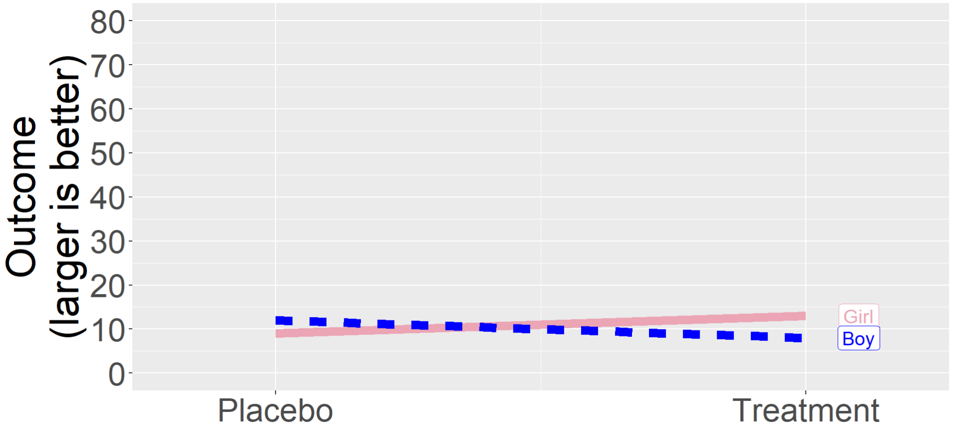
Speaker notes

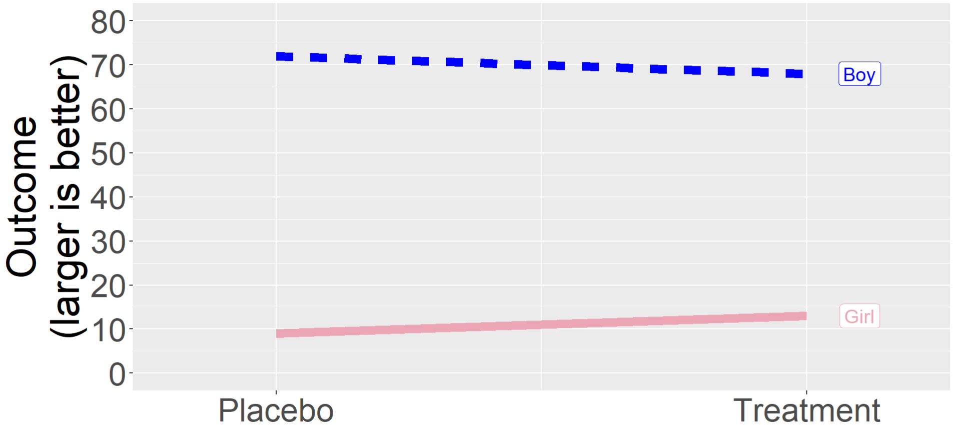
Interactions are important to look for, but if you find one, don’t rejoice. Interactions are a headache. They tell you that a simple interpretation of your research won’t work. That’s important to know, of course, but it also means that you will have to spend more time explaining your results in a paper or presentation.

Interaction plot

- X axis, first categorical variable
- Separate lines for second categorical variable
- Y axis, average outcome

Hypothetical interaction plots, 1 of 4





- No interaction
- Ineffective treatment
- Boys/girls similar

- No interaction
- Ineffective treatment
- Boys fare better than girls

Speaker notes

An interaction plot shows the mean values for each of the two categories. In this example, there is a placebo and a treatment. The outcome is unspecified, but a larger value is presumed to represent a better outcome. This is a pediatric example and the data is subdivided into two populations, boys and girls.

The flatness or steepness of the lines indicates whether patients given the treatment fare better than patients given the placebo.

The separation (if there is any) between the lines measures whether boys fare better or worse than girls.

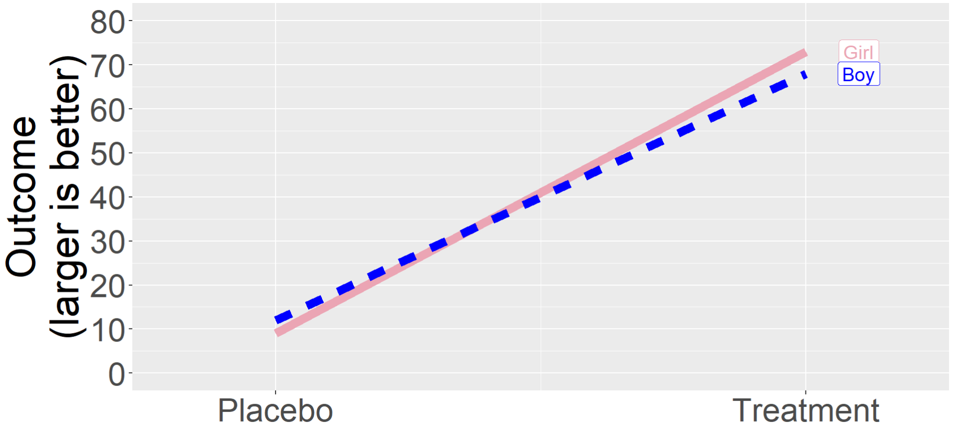
If the lines have roughly the same slope (both are flat or both are steep), then there is no interaction.

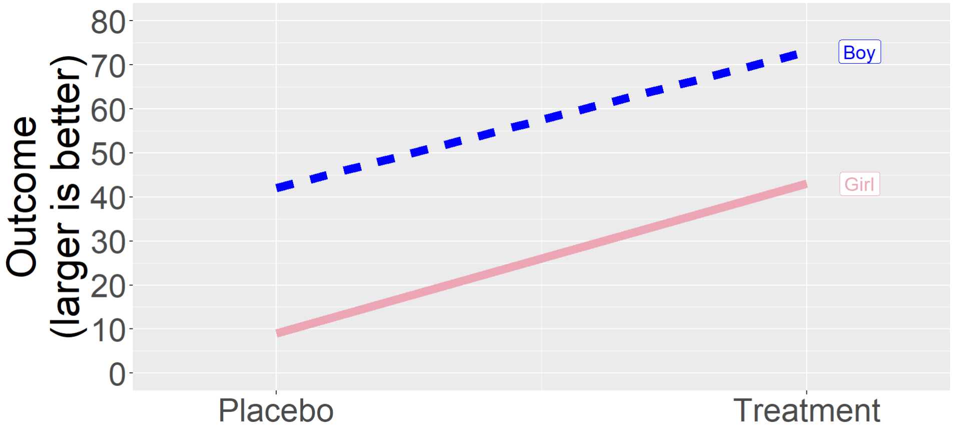
In the plot on the left, the two lines are flat, indicating that the treatment is ineffective. The outcome is not changed from the placebo.

The two lines lie more or less on top of one another. This indicates that there is no difference in average outcome between boys are girls.

In the plot on the right, the two lines are flat. The treatment is ineffective. There is, however, a difference. The average outcome for boys is a lot better both in the placebo group and the treatment group. The lines are roughly parallel, indicating no interaction.

Hypothetical interaction plots, 2 of 4





- No interaction
- Effective treatment
- Boys/girls similar

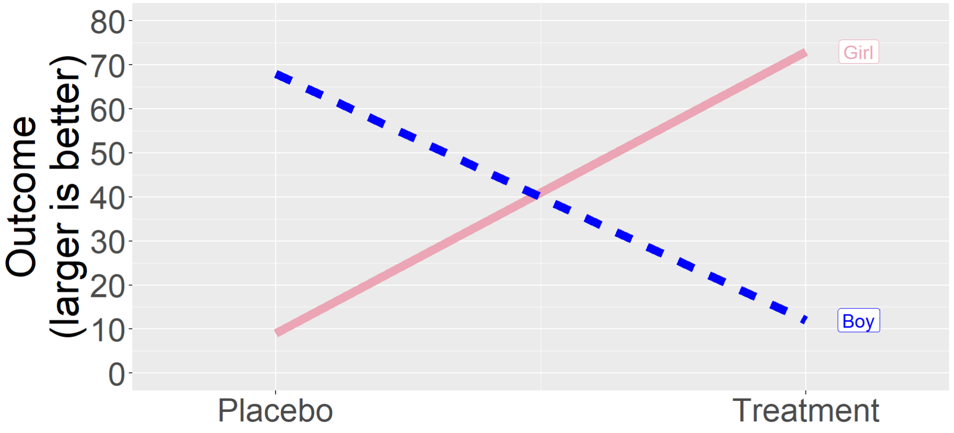
- No interaction
- Effective treatment
- Boys fare better than girls

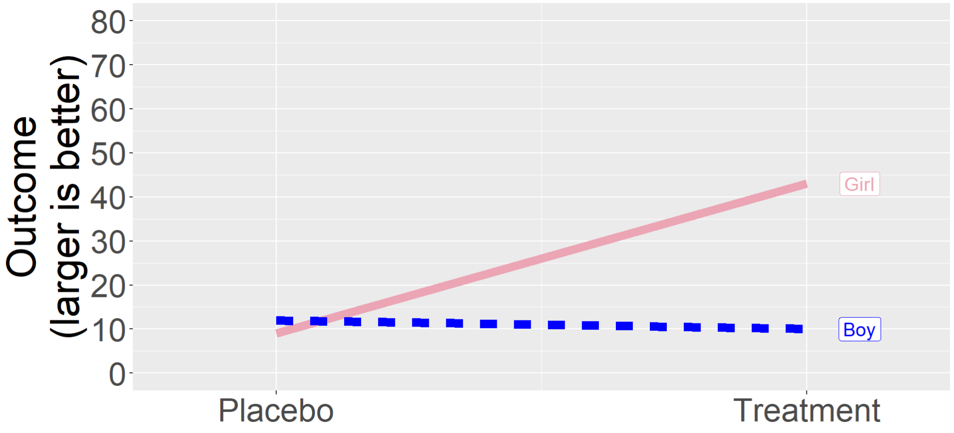
Speaker notes

In the plot on the left, there is a steep slope for both boys and girls. The treatment is effective. There is no separation in the lines. Boys do not fare any better or worse on average than girls.

In the plot on the left, there is a steep slope and a separation between the lines. Boys fare better than girls on average. Both lines have a steep slope. The treatment. The lines are parallel, so there is no interaction.

Hypothetical interaction plots, 3 of 4





- Significant interaction
- Harmful treatment in boys
- Effective treatment in girls

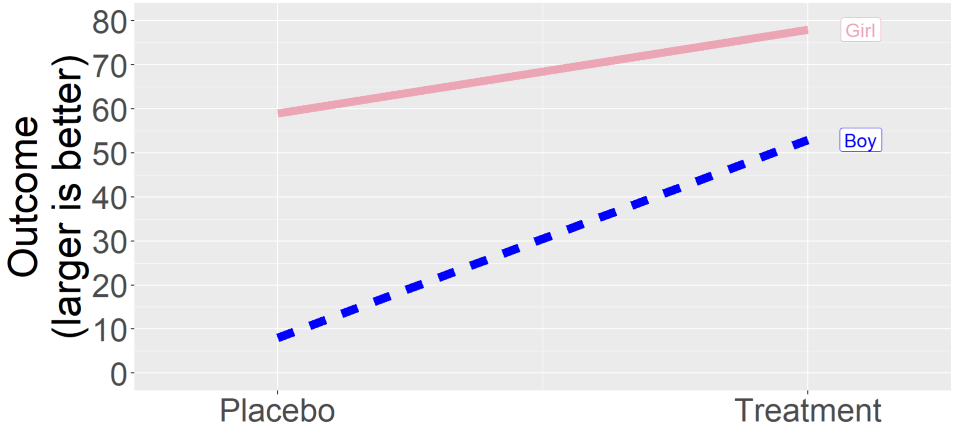
- Significant interaction
- Ineffective treatment in boys
- Effective treatment in girls

Speaker notes

In the plot on the left, the lines are not parallel, so this is evidence of an interaction. In fact, the two lines cross. This is an extreme interaction. Boys fare better on the treatment and girls fare better on the placebo.

In the plot on the right, the lines are not parallel, so this is also evidence of an interaction, but a different sort of interaction. The line for boys is flat and the line for girls is steep. The treatment is worthless for boys, but quite helpful for girls.

Hypothetical interaction plots, 4 of 4

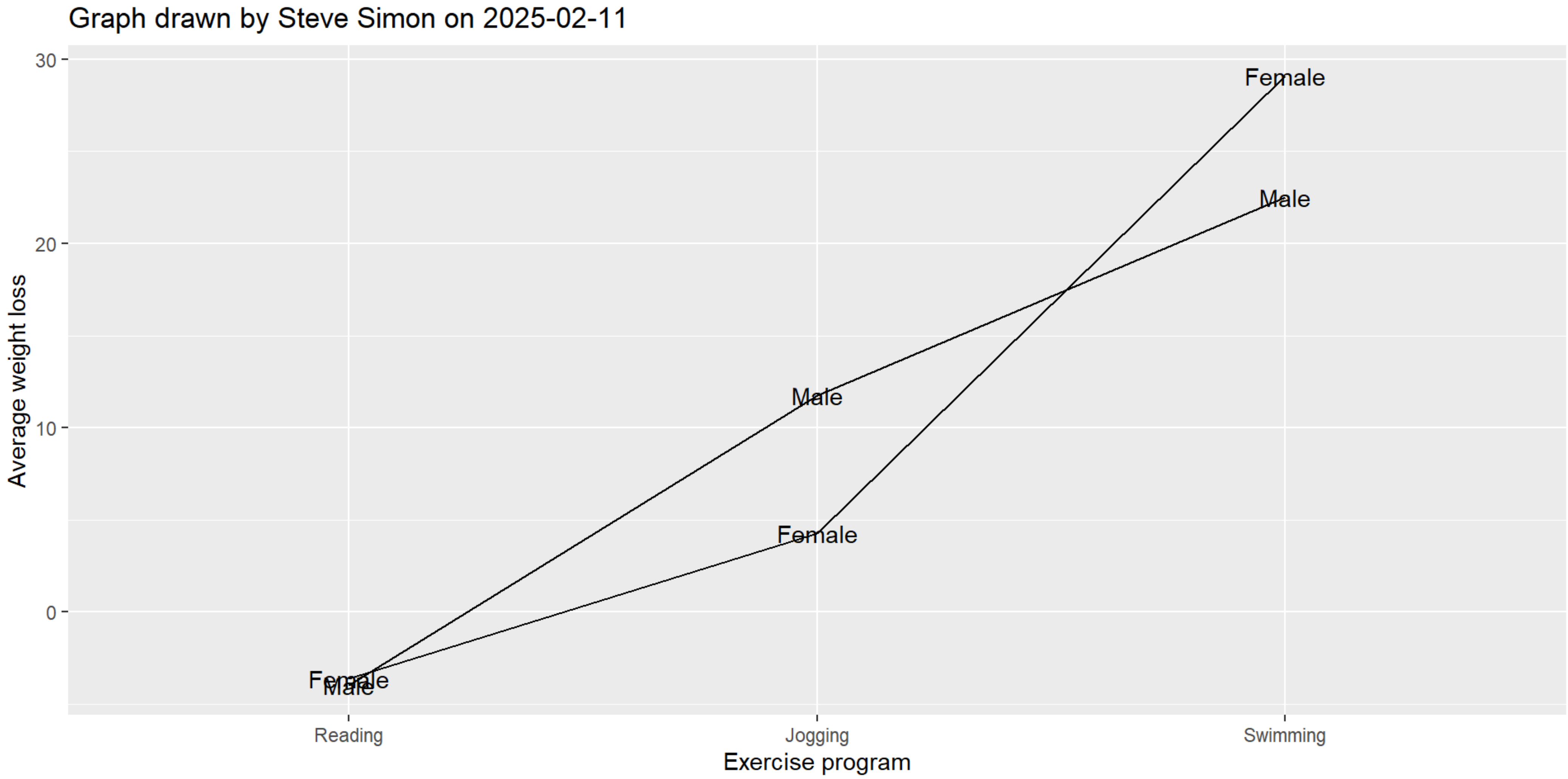


- Significant interaction
- Girls fare better overall
- Effective treatment
- Much more effective in boys

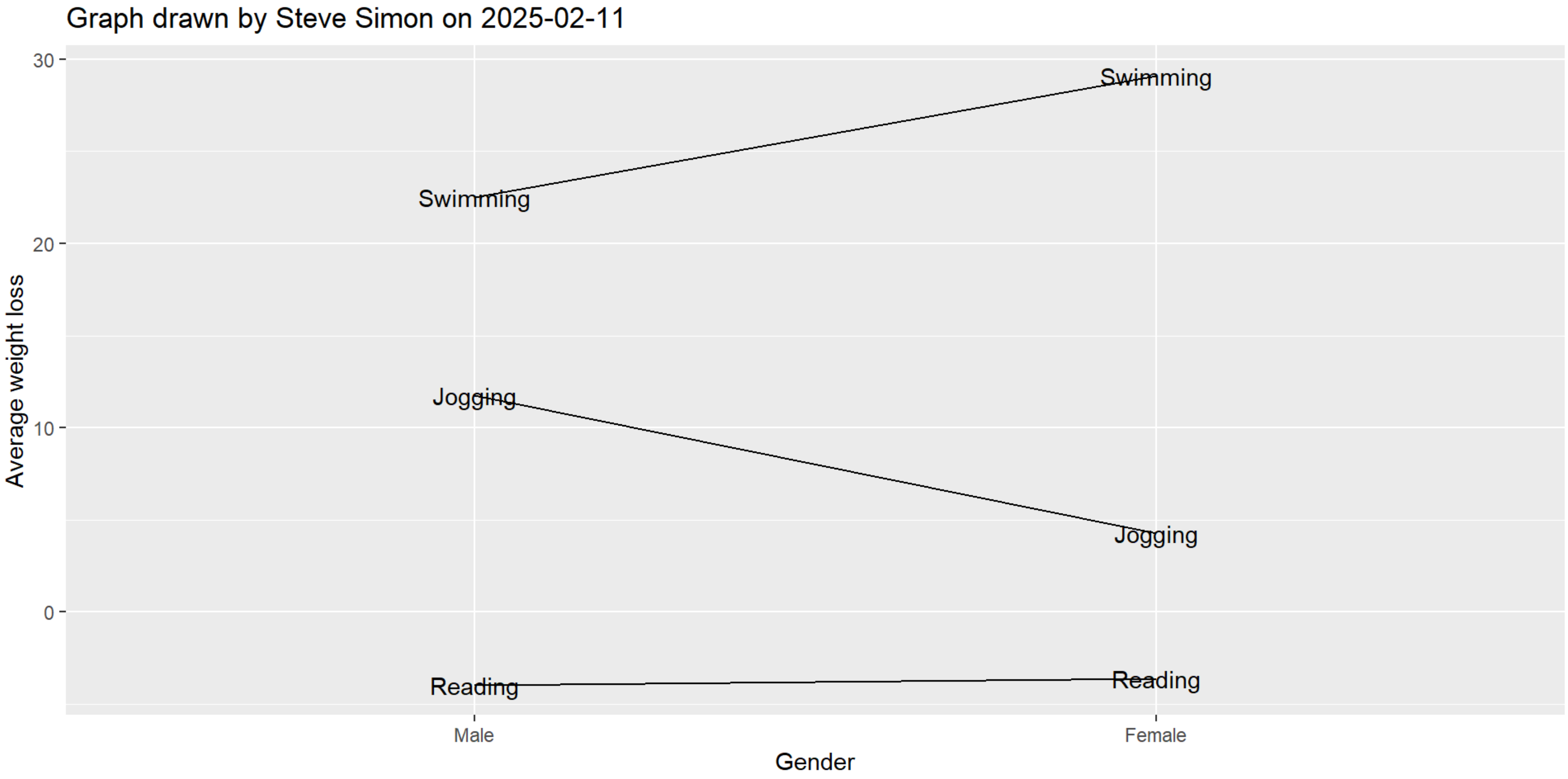
Speaker notes

In this final plot, the lines are not parallel, indicating a third type of interaction. The slope is much steeper for boys. Girls see a moderate improvement on average, but boys see a really large improvement.

Line plot of exercise data



Alternative line plot of exercise data



When you can’t estimate an interaction

- Special case, n=1
  - Only one observation for categorical combination

Speaker notes

There is a special case where you have two categorical independent variables and you cannot estimate an interaction. If you have n=1, exactly one observation for each combination of your two categorical variables, then you don’t have enough degrees of freedom to estimate an interaction and still be able to test whether that interaction is statistically significant.

It’s sort of like that old joke I told about married life (it’s okay but you lose a degree of freedom). Interactions cause an even bigger loss of degrees of freedom and in the case with only one observation per combination of categories, you lose enough degrees of freedom that it is not marriage, it being in prison.

Live demo, Interactions in analysis of variance

Live demonstration of part 1 of simon-5502-04-demo.qmd

Summary

- What you have learned
    - Review one factor analysis of variance
    - Multiple factor analysis of variance
    - Checking assumptions of analysis of variance
    - Interactions in analysis of variance