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.

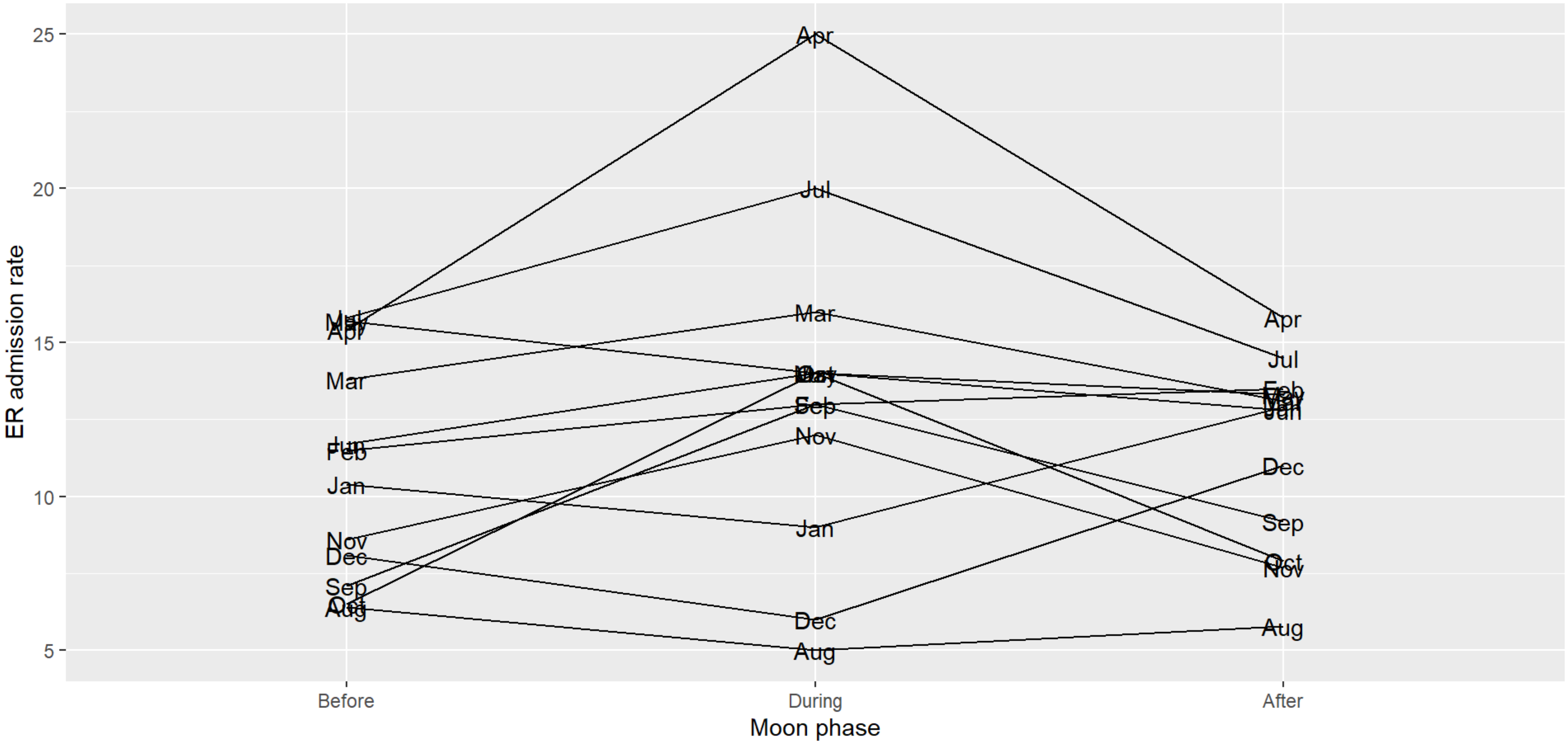
Full moon data

- 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



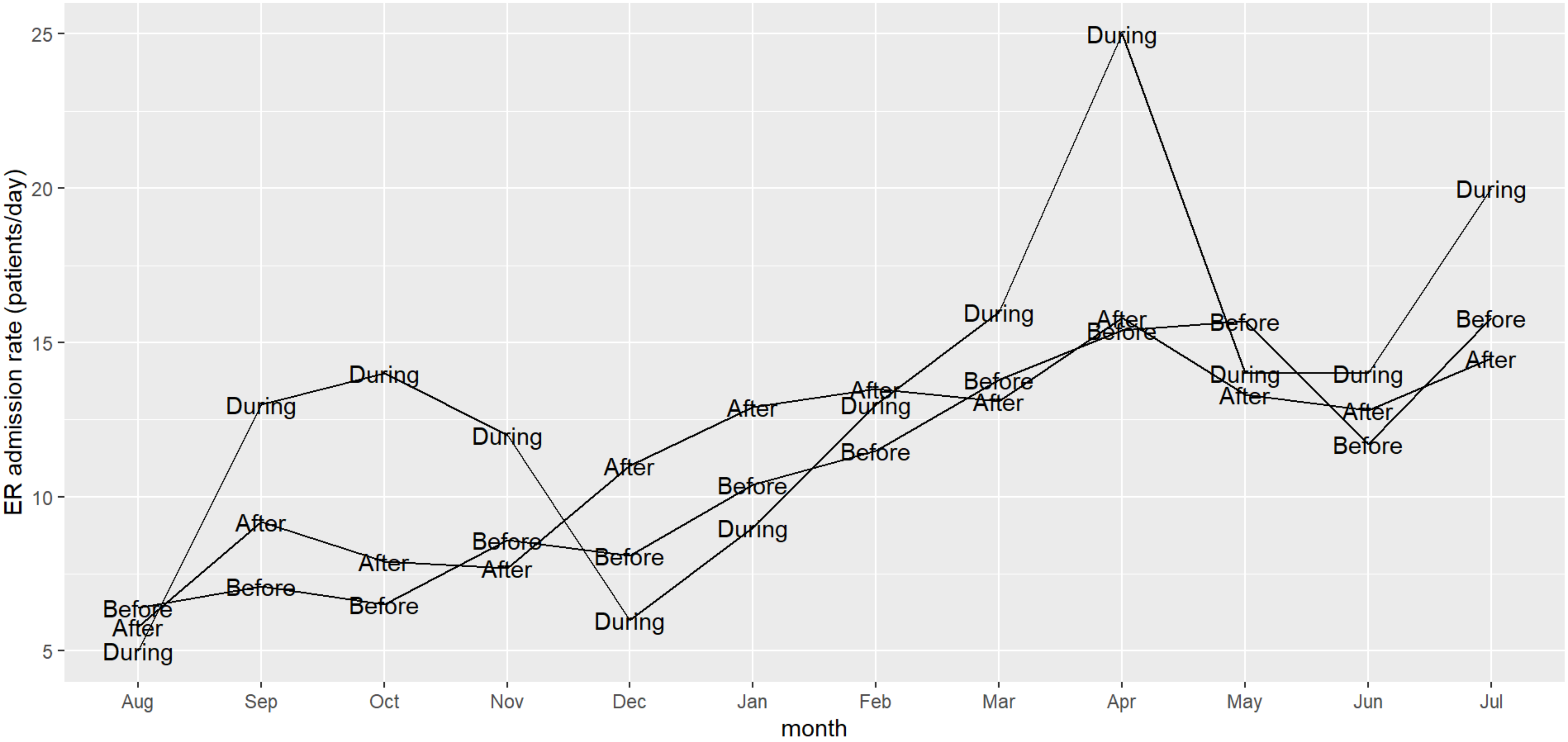
Steve Simon, 2025-02-08, CC0

Speaker notes

This plot shows a separate line for each month and tracks the admission rates before, during, and after a full moon.

This plot is very busy and hard to interpret. The pattern that you might expect to see is an inverted V for each month, showing that ER admissions jump during a full moon and fall back down again afterwards. This seems to hold for some months, but not others. So if there really is an effect of the full moon on admissions, it is not very strong or consistent.

Alternative line graph of full moon data



Steve Simon, 2025-02-08, CC0

Speaker notes

You can draw this plot differently with three lines, one for each phase and show how the ER admission rates change over the months. There appears to be a general upward trend over time. The before, during, and after lines criss-cross quite often indicating that there may not be a consistent effect of phase of the moon. But it does look like there are more months where the during line rises well above the other lines.

Descriptive statistics by moon phase

```
# 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
```

Speaker notes

The F-ratio (1.17) is close to zero and the p-value (0.322) is large. There is no evidence that the average admission rates change by phase of the moon.

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
```

Speaker notes

The reference level was set at moon=Before. So the intercept represents the estimated average admission rate before a full moon.

The first slope is interpreted as the estimated average change in admission rates when the phase of the moon shifts from before a full moon to during a full moon. There is an increase of 2.5 patients per day, but this is not statistically significant. There is too much noise in the data.

The second slope is interpreted as the estimated average difference between the admission rates after a full moon compared to before a full moon. This change is very small (0.5 patients per day) and is also not statistically significant.

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.

Testing the global hypothesis

- $H_0 : \ \alpha_i = 0, \ \beta_j = 0$ for all i,j
- H_1: At least one α or $\beta \neq 0$

```
Model    rss         sumsq
null     SSE(null)  NA
full     SSE(full)  SSR(full)
                     = SSE(null)-SSE(full)

Note: SSE(null) is usually called SST or SS_Total.

Degrees of freedom = n-a-b+1, n-1, a+b-2
```

Speaker notes

To test the combined effect of both factors, you are testing the null hypothesis that all the alphas and betas are equal to zero versus the alternative hypothesis that one of more of the alphas and betas are not equal to zero.

You test this hypothesis by comparing the unexplained variation in the null model, SSE(null) to the unexplained variation in the full model, SSE(full). The difference between these is the amount of variation that this model explains, SSR(full).

The more traditional layout

```
            SS       df    MS
Regression  SSR    a+b-2   MSR = SSR/df
Error       SSE   n-a-b+1  MSE = SSE/df
Total       SST     n-1

F = MSR / MSE

SSR = SSR(full)
SSE = SSE(full)
SST = SSE(null)
```

Speaker notes

Here is the more traditional layout. This reinforces the additive nature of these models. Explained error (regression) plus unexplained error equals total.

You may see a different term than “Regression” such as “Model”.

The F-ratio is a measure of how much evidence we have that the two factors help in predicting the outcome. If it is close to 1, you should accept the null hypothesis. If it is a lot larger than 1, you should reject the null hypothesis.

R-squared

- $R^2 = \frac{explained\ variation}{total\ variation} = \frac{SSR(full)}{SSE(null)}$
- $R^2 = 1 - \frac{unexplained\ variation}{total\ variation} = 1 - \frac{SSE(full)}{SSE(null)}$

Speaker notes

R-squared is defined in the multiple factor analysis of variance pretty much the same way as it is in multiple linear regression. It is the ratio of explained variation to total variation. Remember that total variation is the variation under the null model

Testing the partial hypothesis

- $H_0 : \ \beta_j = 0$ for all j
- H_1: At least one $\beta \neq 0$

```
Model    rss            sumsq
partial  SSE(partial)   NA
full     SSE(full)      SSR(full|partial)
                          = SSE(partial)-SSE(full)

Degrees of freedom = n-a-b+1, n-1, a+b-2
```

Speaker notes

You can test the partial hypothesis, which is the effect of the second factor represented by the betas in a model that already includes the first factor. To do this, you note the sum of squares error for the partial model, the one with only the first factor to the sum of squares error for the full model, the one that includes both the first and the second factor. The difference between these is the partial sum of squares for regression. This partial sum of squares for regression measures the additional variation explained by the second factor after whatever the first factor accounted for.

Partial R-squared

- $partial\ R^2 = \frac{partial\ explained\ variation}{total\ variation} = \frac{SSR(full|partial)}{SSE(null)} = \frac{S}{\ }$

Speaker notes

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
```

Speaker notes

There are reference levels for both moon and month. For moon, the reference label is before. For month, the reference label is Aug. The moonDuring slope is the estimated average change in admission rates when switching from before a full moon to after a full moon holding month constant The moonAfter slope is the estimated average change in admission rates when comparing after a full moon to before a full moon.

The estimates for month do not all fit on this slide, but monthSep represents the estimated average change in admissions when you move from the reference level (August) to September, holding phase of the moon constant. The slope for month Oct represents the estimated average change in admission rates fwhen yopu move from August to October, holding phase of the moon constant. The remaining slope terms have similar interpretations.

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

The F-ratio (7.13) is large and the p-value is small. There is a statistically significant effect in at least some of the factors. It could be differences among the months or differences among the phases or differences among both. This is a global test, so you do not know which months differ (if any) or which phases differ (again, if any).

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
```

Speaker notes

You compare the moon model with the moon plus month model by looking at the sum of squared error. In R, this is labeled as rss (residual sum of squares). The moon model has a large amount of unexplained variation (583) while the moon plus month model has a much smaller amount of unexplained variation (128). The large F ratio and the small p-value indicates that there is a statistically significant improvement in prediction when you use moon and month over a model using just moon alone.

R-squared values

```
# A tibble: 3 × 3
  model      r.squared deviance
  <chr>          <dbl>    <dbl>
1 Null          0          625.
2 Moon          0.0664     583.
3 Moon Month    0.795      128.
```

Speaker notes

The null model has a large amount of unexplained variation (625). The moon model can account for about 7% of this variation, while the moon plus month model accounts for almost 80% of the variation.

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

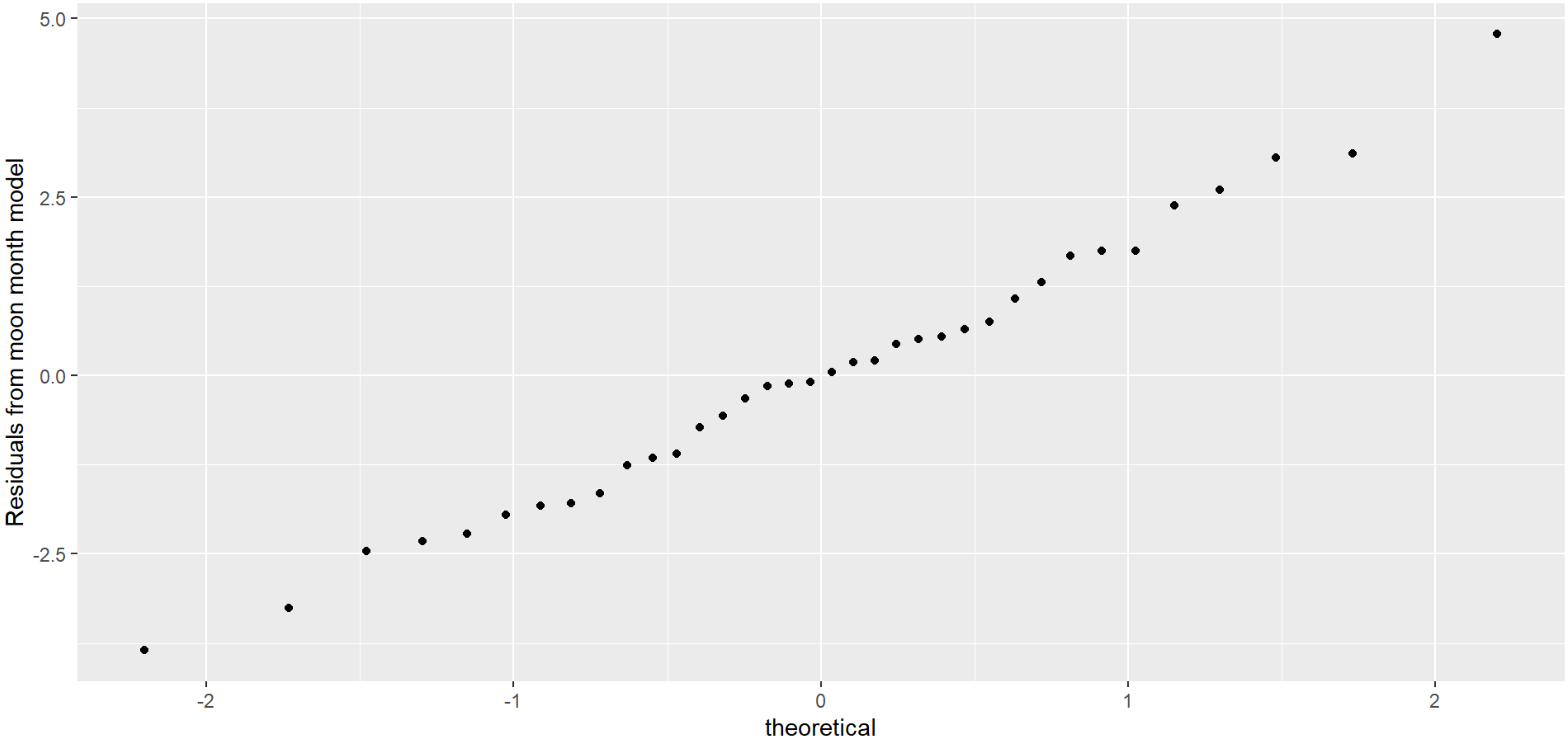
Assumptions

- 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

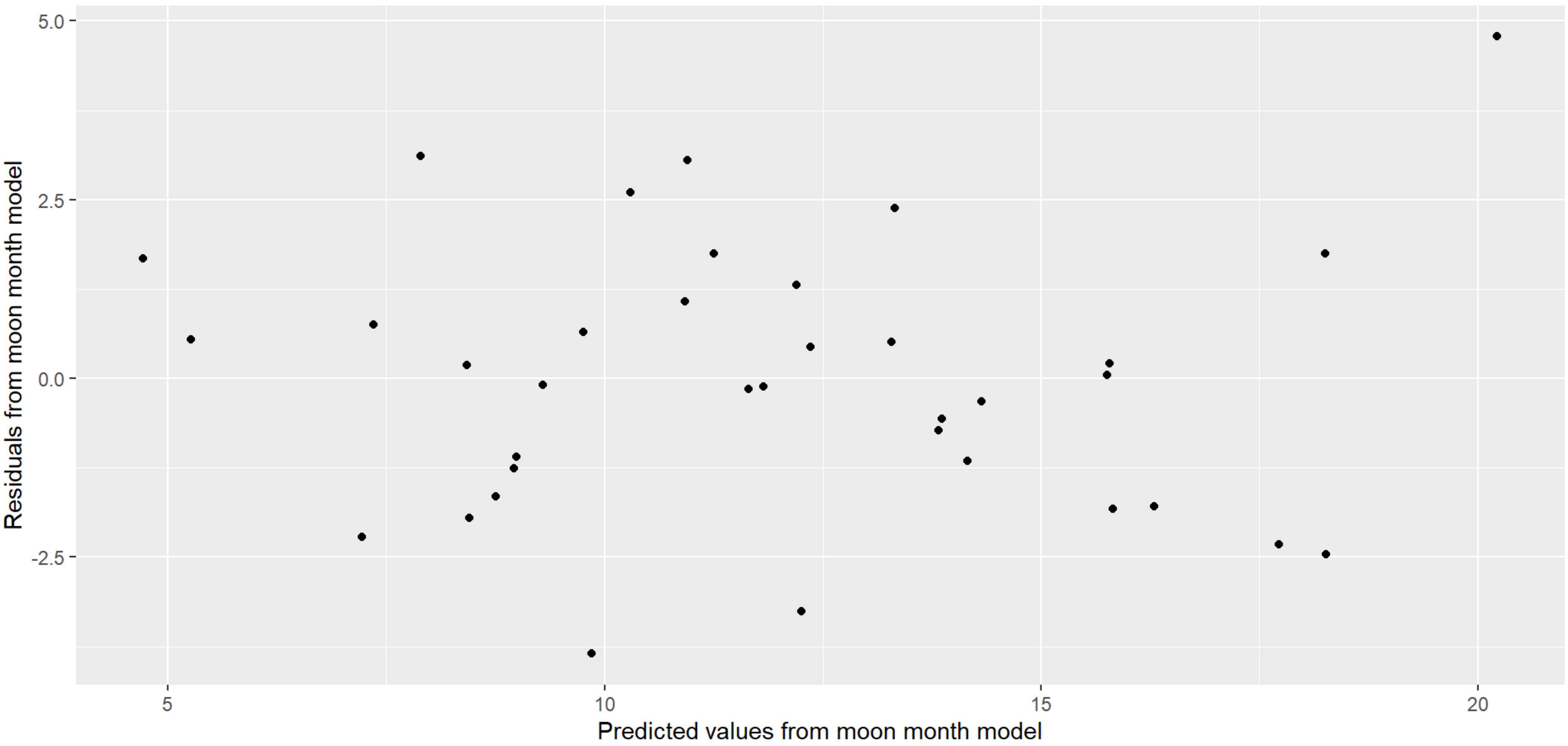


Steve Simon, 2025-02-11, CC0

Speaker notes

For the moon analysis, the normality assumption appears to be satisfied.

Residual versus predicted value plot



Steve Simon, 2025-02-11, CC0

Speaker notes

The homogeneity assumption also appears to be satisfied. A violation of homogeneity is often a fanning out patter where you see more variation for large predicted values than for small predicted values.

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

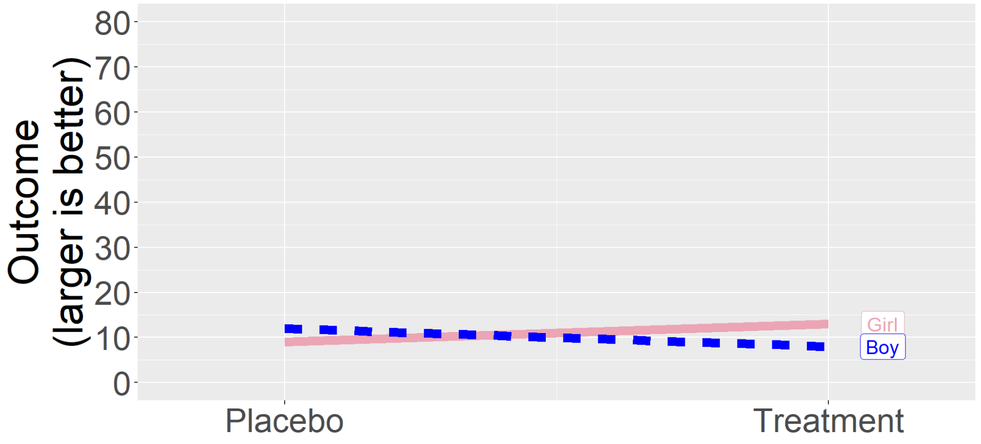
Speaker notes

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.

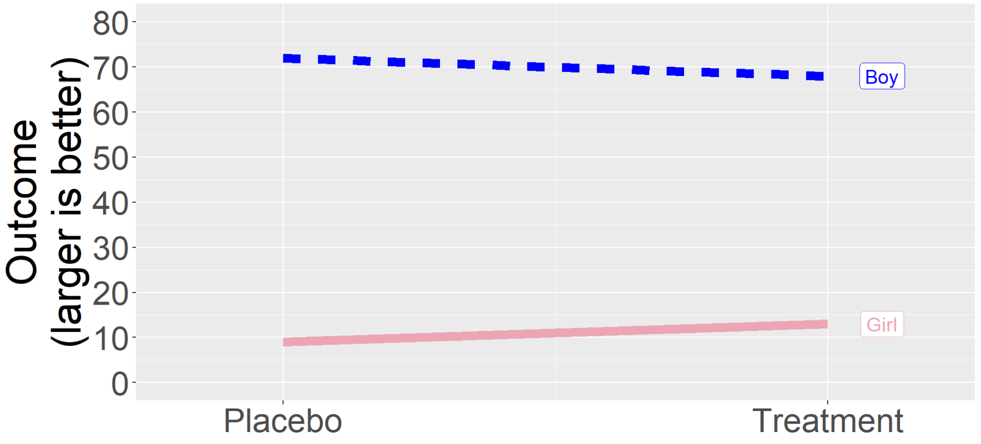
Interaction plot

- 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.

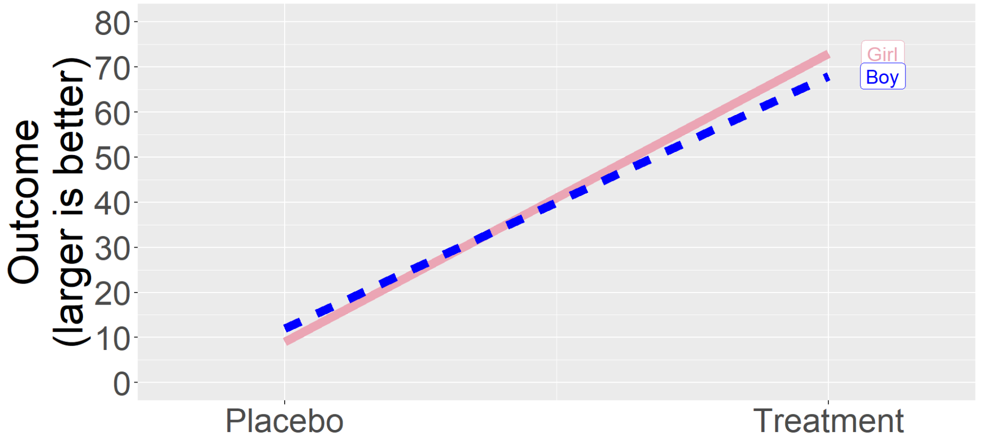
If the lines have roughly the same slope (both are flat or both are steep), then there is no interaction.

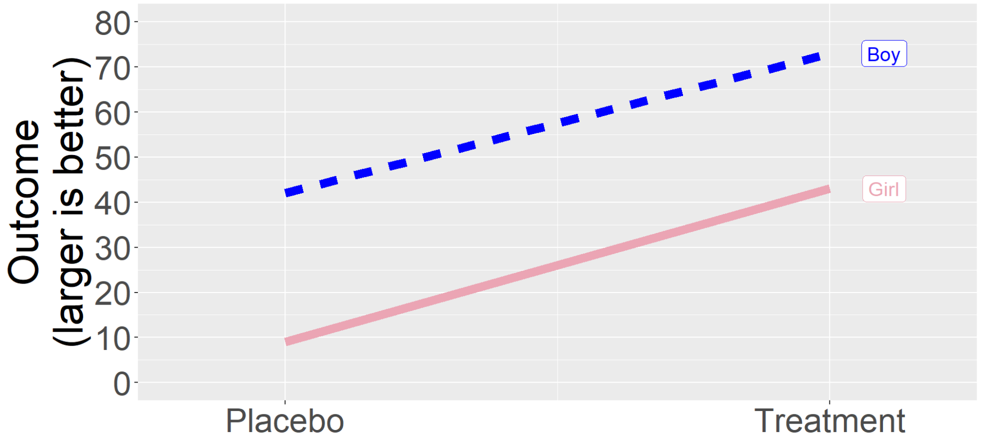
In the plot on the left, the two lines are flat, indicating that the treatment is ineffective. The outcome is not changed from the placebo.

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

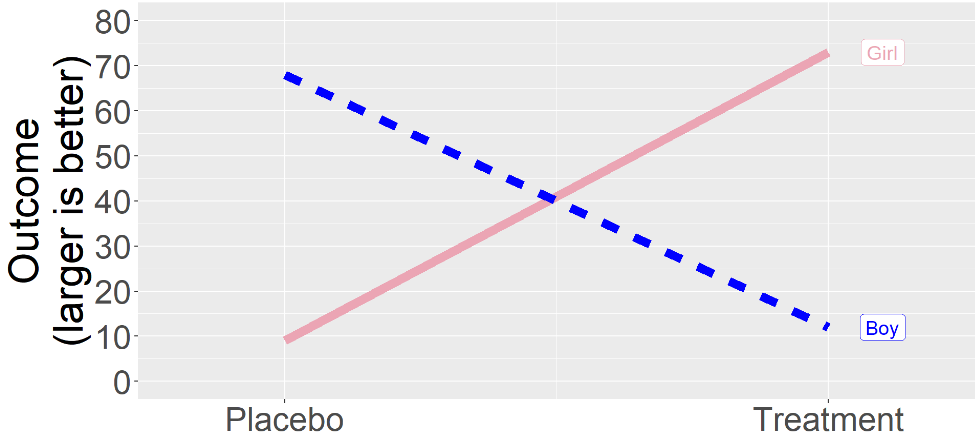
- No interaction
- Effective treatment
- Boys fare better than girls

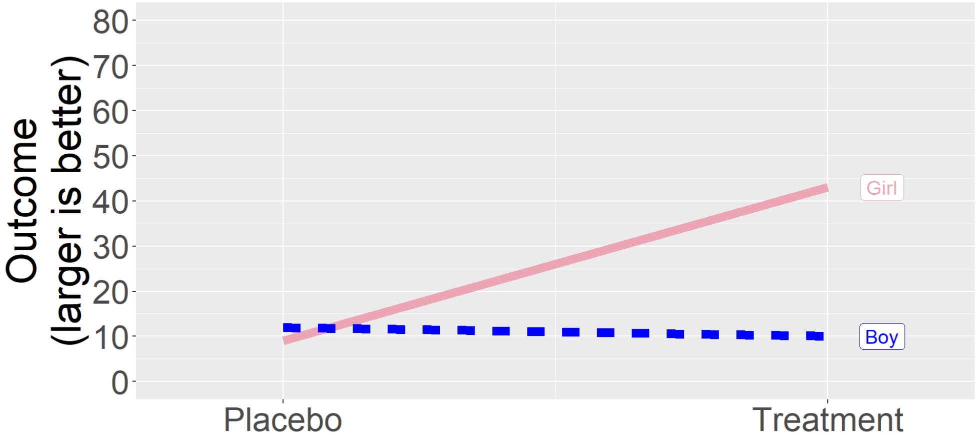
Speaker notes

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

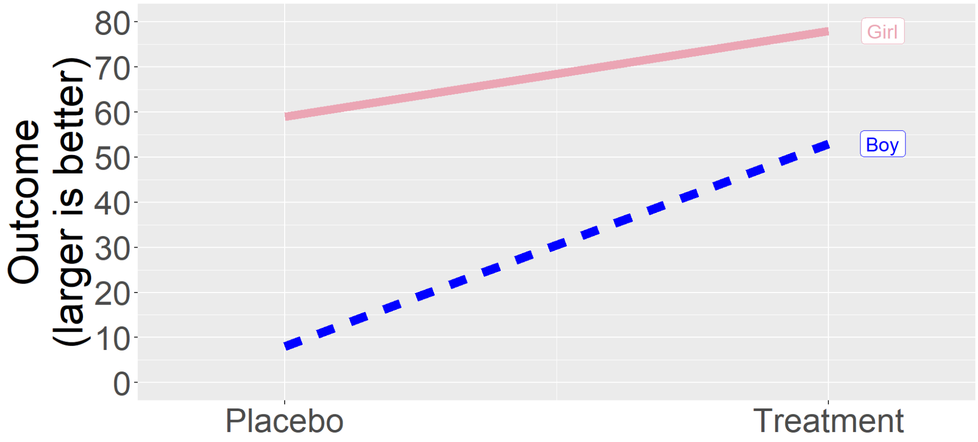
- Significant interaction
- Ineffective treatment in boys
- Effective treatment in girls

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.

Group means for tracking data

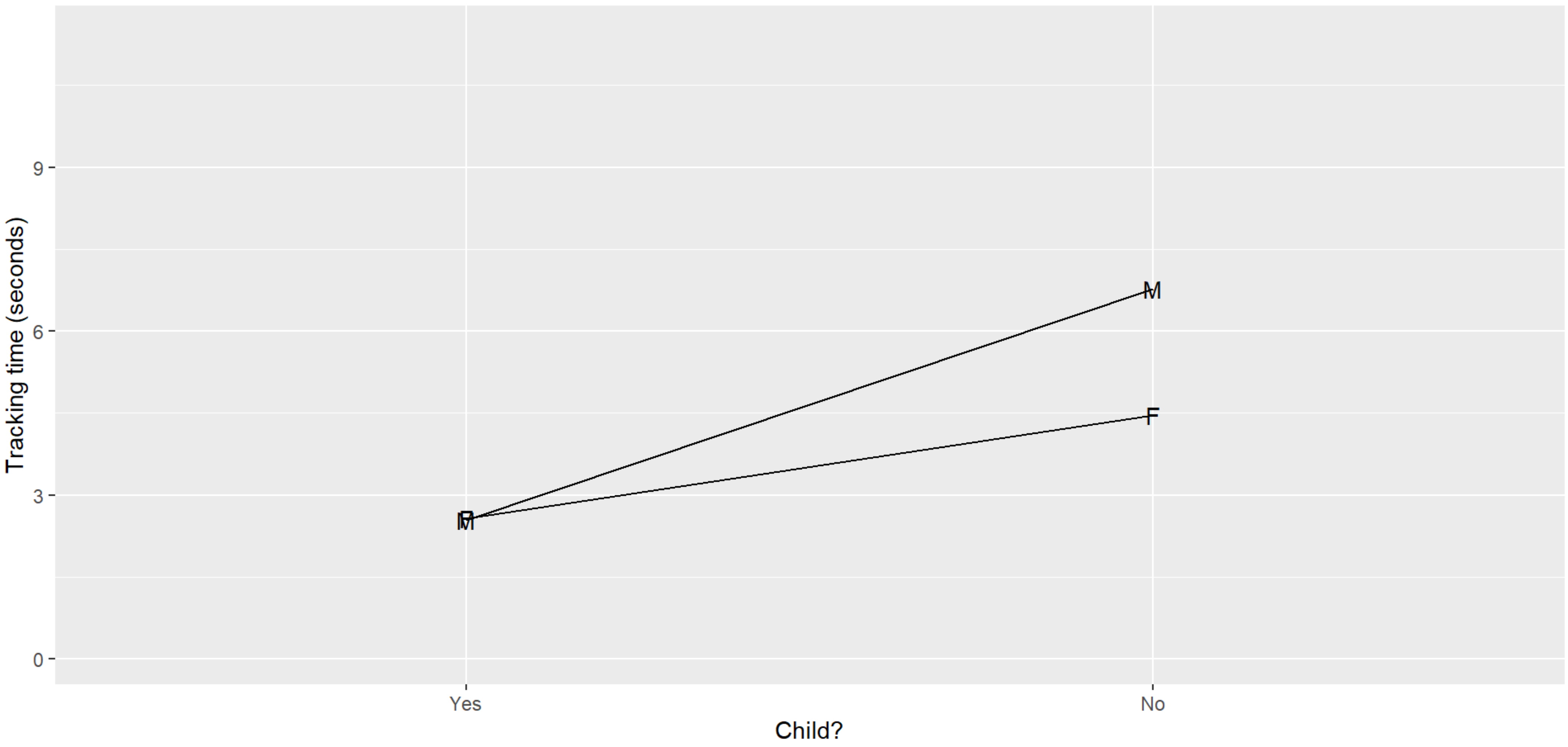
```
# A tibble: 4 × 3
  child gender mean_trial4
  <fct> <chr>        <dbl>
1 Yes   F             2.57
2 Yes   M             2.54
3 No    F             4.46
4 No    M             6.77
```

Speaker notes

The data used here is an exercise in motor skills where a subject is asked to maintain contact with a small area on a rotating disk. There were both male and female subjects and a wide range of ages. I decided to classify age into child=Yes (18 and younger) and child=NO (over 18).

The means for this study are interesting. There is almost no difference between male and female children. Adults do better than children, but the gap is much larger when comparing male adults to male children.

Line plot of tracking data

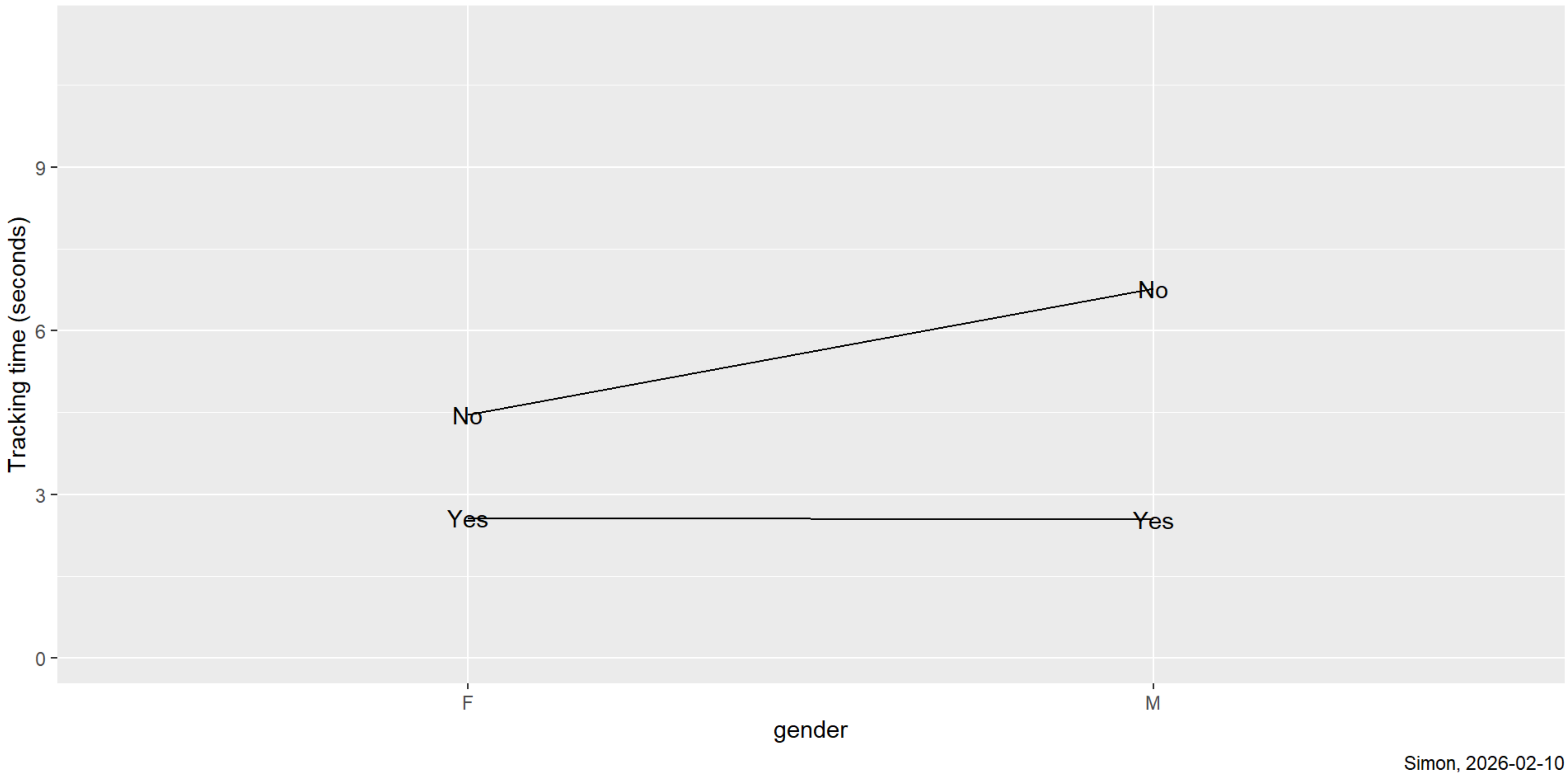


Simon, 2026-02-05

Speaker notes

The two lines are right on top of one another for child=Yes and diverge for child=No. So there is a gender gap, but only for adults.

Alternative line plot of exercise data



Speaker notes

This is an alternate display of the interaction. The chld=Yes line is flat indicating no change due to gender. The child=No line is above the child=Yes line, indicating superior performance for adults versus children. The gap is wider for males than for females.

You should try both graphs and then display only one. Display the one that portrays the clearest picture of what is going on.

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