MEDB 5501, Module12

2025-11-11

Topics to be covered

- What you will learn
  - Two factor analysis of variance
  - Relationship to linear regression
  - Checking assumptions
  - R code for two factor analysis of variance
  - Interactions
  - R code for interactions
  - Your homework

Two factor analysis of variance

- Continuous outcome
- Two categorical predictors
- Example
  - Hearing test (decibels at high frequency)
  - Age group (Old or Young)
  - Gender (Female or Male)

Speaker notes

Two factor analysis of variance uses two categorical variables to predict a continuous outcome. We will focus today on the balanced case. In the balanced case, each combination of category levels has the same number of observations or if they differ, they differ in a proportional way.

Balanced data

- Proportional number in each category level combination group
  - 3 old females, 3 old males, 3 young females, 3 young males
  - 6 old females, 6 old males, 2 young females, 2 young males

Speaker notes

The first case each combination of gender and age has exactly three observations. The second case is also balanced. The ratio of old to young is 3 to 1 both for females and for males.

Unbalanced data

- Unequal numbers in some category combinations
  - 3 old females, 3 old males, 3 young females, 2 young males
- Extreme case: empty category combinations
  - 3 old females, 3 old males, 3 young females, 0 young males

Speaker notes

In the unbalanced case the observations are not equal or proportional. These cases are quite complex from several perspectives.

The root cause of all the complexity is that there is more than one way to calculate averages.

Historical perspective on unbalanced data, 1

Speaker notes

Let me take a few minutes to discuss some of the people behind the work on unbalanced data. I don’t want you to delve too deeply into this, as the work is highly mathematical. But I do have a special fondness for this work because I know some of the people who were pioneers in this area.

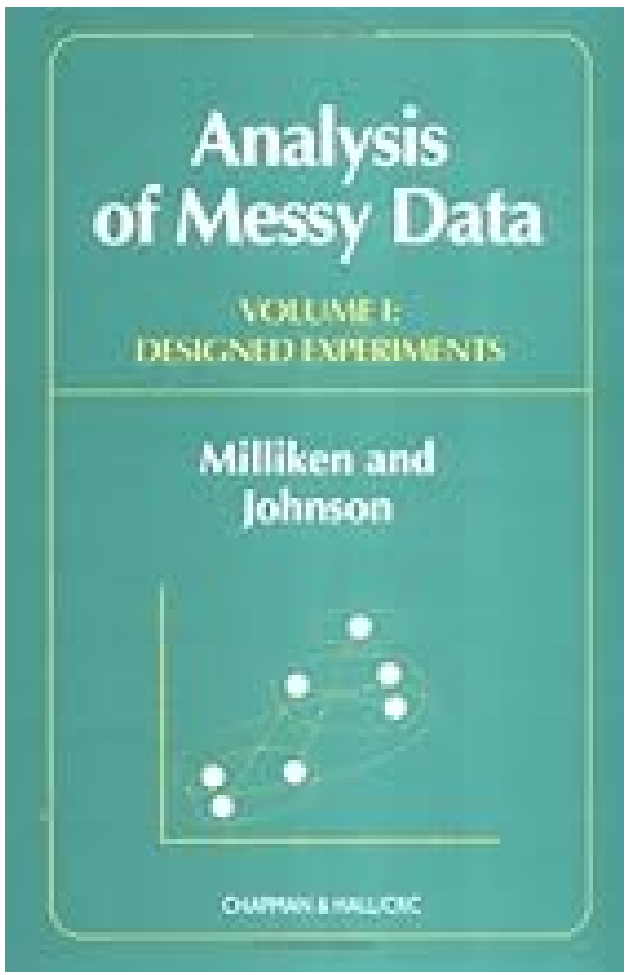
One of the pioneers in documenting the complexities of unbalanced data is George Milliken, a Statistics faculty member (long since retired) from Kansas State University.

Historical perspective on unbalanced data, 2

Speaker notes

A second pioneer in this area is Dallas Johnson. He is also a faculty member in Statistics at Kansas State University and also long retired.

Historical perspective on unbalanced data, 3



Speaker notes

Together, they wrote a series of books with the provocative title, Analysis of Messy Data. This is the first volume in the series, published in 1984. The book deals largely with unbalanced data in analysis of variance. I must admire the bravery of writing this book. At a time when I avoided the complexities of unbalanced data in my teaching, Milliken and Johnson dove straight in.

I've met both Milliken and Johnson at several "Agriculture and Statistics" conferences held for many years at Kansas State University. I've always been more of the health care side than the agricultural side of statistics, but there is a lot that I have learned from these conferences. An interesting sidelight is that the conference always held line dancing lessons on the Sunday before the conference start. I and other geeky statisticians would get up and awkwardly follow the instructions. I never got very good at line dancing, but it was always a lot of fun.

Historical perspective on unbalanced data, 4

Speaker notes

A third and more receint pioneer in this area is Russell Lenth, a faculty member in Statistics at the University of Iowa and also long retired. I know Russ very well, as he was my dissertation adviser when I got my PhD from Iowa in 1982. He wrote the definitive package in R for unbalanced data, emmeans.

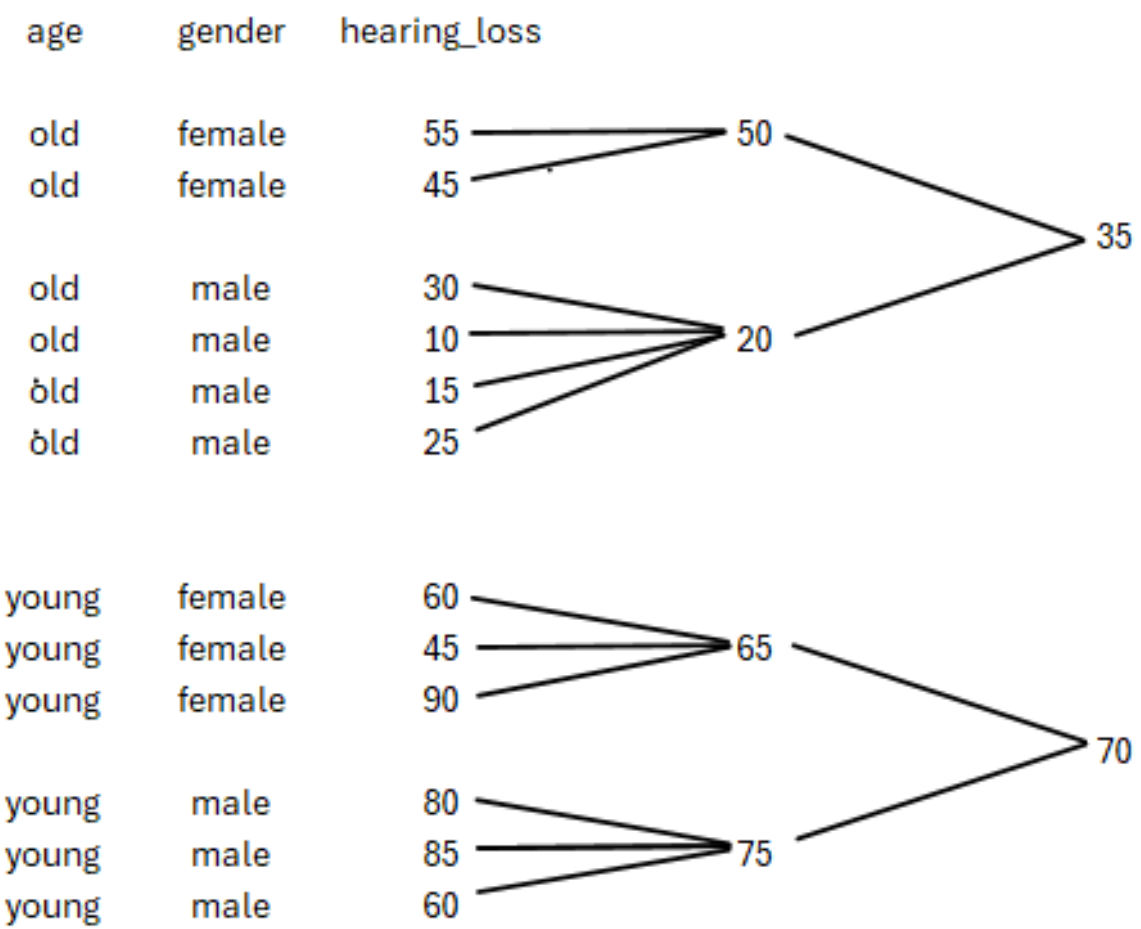
A simple illustration of the complexities of unbalanced data, 1

age	gender	hearing_loss
old	female	55
old	female	45
old	male	30
old	male	10
òld	male	15
òld	male	25
young	female	60
young	female	45
young	female	90
young	male	80
young	male	85
young	male	60

Speaker notes

Here is some hypothetical data to illustrate two different ways of computing average hearing loss in old patients.

A simple illustration of the complexities of unbalanced data, 2



Speaker notes

To get the average for all old patients, start by computing the average for old female patients (50) and old male patients (20). Then average the old female mean and the old male mean to get an old patient mean of 35.

A simple illustration of the complexities of unbalanced data, 3

age	gender	hearing_loss	
old	female	55	
old	female	45	
old	male	30	30
old	male	10	
old	male	15	
old	male	25	
young	female	60	
young	female	45	
young	female	90	70
young	male	80	
young	male	85	
young	male	60	

Speaker notes

You could also compute the average across all six old patients in a single step. This produces an old patient mean of 30.

Why the difference? In the single step calculation the four old males carry more weight than the two old females. In the two step calculation, old males and old females get equal weight.

You might argue that the equal weighting makes more sense, and in some contexts it does. You might consider this an effort to adjust for the imbalance.

I won’t go into more detail about this, but hope to elaborate on the topic in Biostats II.

Mathematical model, 1

- Y_{ijk}
  - i = which level of first category
  - j = which level of second category
  - k = which patient within a category combination

Speaker notes

With two levels, you need three subscripts (i, j, k) to keep track of the observations.

Mathematical model, 2

Age	$Gender$	$Outcome$
Old	$Female$	Y_{111}
Old	$Female$	Y_{112}
Old	$Female$	Y_{113}
Old	$Male$	Y_{121}
Old	$Male$	Y_{122}
Old	$Male$	Y_{123}
$Young$	$Female$	Y_{211}
$Young$	$Female$	Y_{212}
$Young$	$Female$	Y_{213}
$Young$	$Male$	Y_{221}
$Young$	$Male$	Y_{222}
$Young$	$Male$	Y_{223}

Speaker notes

Here’s a simple example where the first categorical predictor and the second categorical predictor have two levels and there are three observation in each combination.

Notice that unlike Milliken and Johnson, I am avoiding the complexities of imbalance in this example.

Mathematical model, 3

- $Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$
  - $i = 1, \ldots, a,\ j = 1, \ldots, b$
  - $\Sigma\alpha_i = 0,\ \Sigma\beta_j = 0$
  - ϵ_{ijk} is $N(0, \sigma)$
- $\bar{Y}_{i..}$ is the average for the ith level of first factor
- $\bar{Y}_{.j.}$ is the average for the jth level of second factor
- $\bar{Y}_{...}$ is the average for all of the data

Speaker notes

The mathematical model includes an overall mean (μ), a deviation from the overall mean associated with the different levels of the first factor (α_i), a deviation from the overall mean associated with the different levels of the second factor (β_j) and an error term (ϵ_{ijk}).

Because the alphas and betas are deviations, they have to sum to zero. The error term is assumed to be normally distributed and the standard deviation is the same for all the data.

You will need to compute averages for the first factor, $\bar{Y}_{i\cdot\cdot}$, the second factor, $\bar{Y}_{\cdot j\cdot}$, and an overall mean, $\bar{Y}_{\cdot\cdot\cdot}$ which is an average across all of the data.

Mathematical model, 4

- $SS(Total) = \Sigma_i \Sigma_j \Sigma_k \ (Y_{ijk} - \bar{Y}_{...})^2$
  - df=abn-1
- $SS(A) = \Sigma_i \ bn(\bar{Y}_{i..} - \bar{Y}_{...})^2$
  - df=a-1
- $SS(B) = \Sigma_j \ an(\bar{Y}_{.j.} - \bar{Y}_{...})^2$
  - df=b-1
- $SS(Error) = SS(Total) - SS(A) - SS(B)$
  - df=(abn-1)-(a-1)-(b-1)

Speaker notes

To assess the impact of the two categorical predictors, you compute sums of squares. SS(Total) represents the deviation of the individual values from the overall mean. SS(A) represents deviations of the first category means from the overall mean. SS(B) reprsents deviations of the second category means from the overall mean. Whatever is left over is SS(Error).

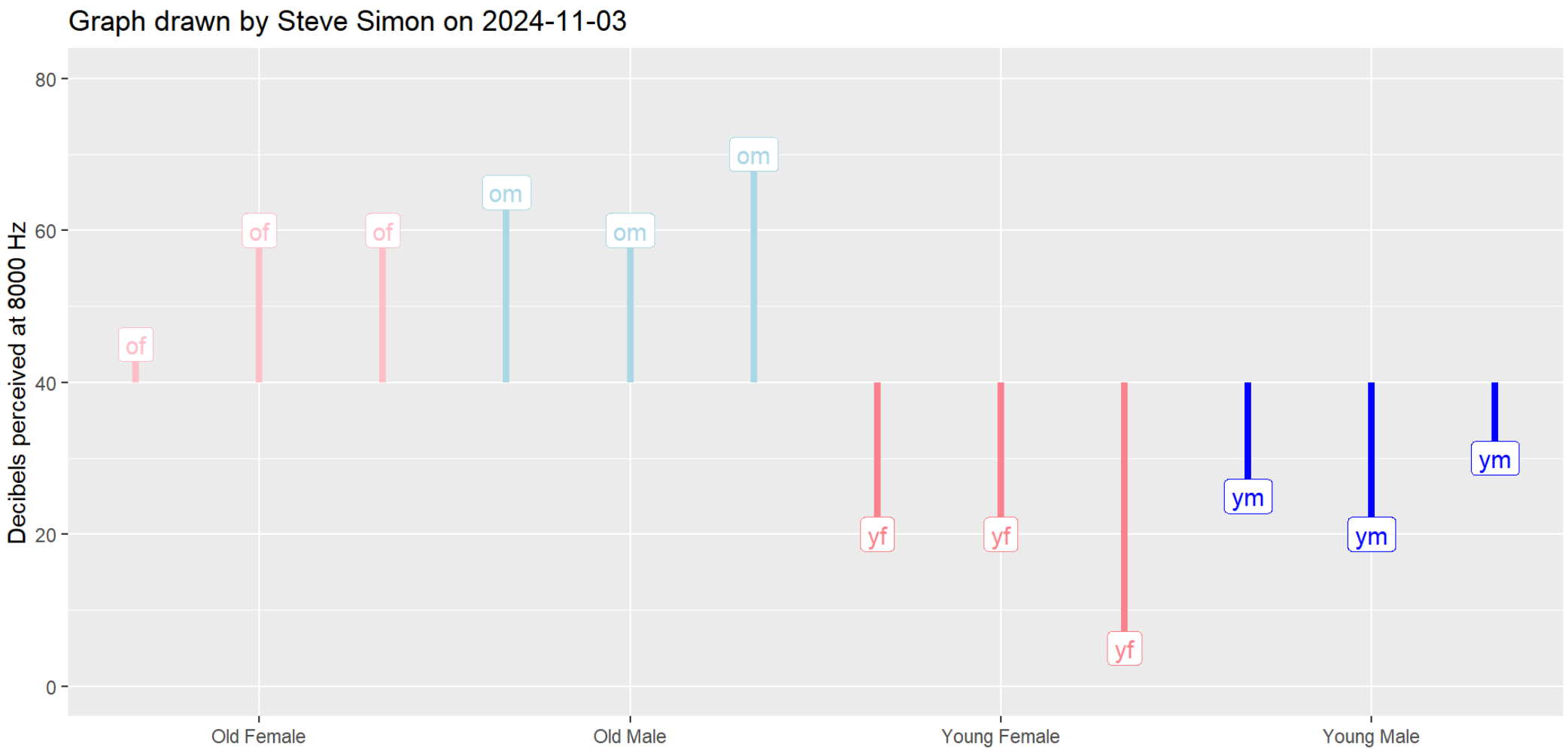
Artificial data

```
# A tibble: 12 × 5
      id age   gender code     db
   <int> <chr> <chr>  <chr> <dbl>
 1     1 old   female of       45
 2     2 old   female of       60
 3     3 old   female of       60
 4     4 old   male   om       65
 5     5 old   male   om       60
 6     6 old   male   om       70
 7     7 young female yf       20
 8     8 young female yf       20
 9     9 young female yf        5
10    10 young male   ym       25
11    11 young male   ym       20
```

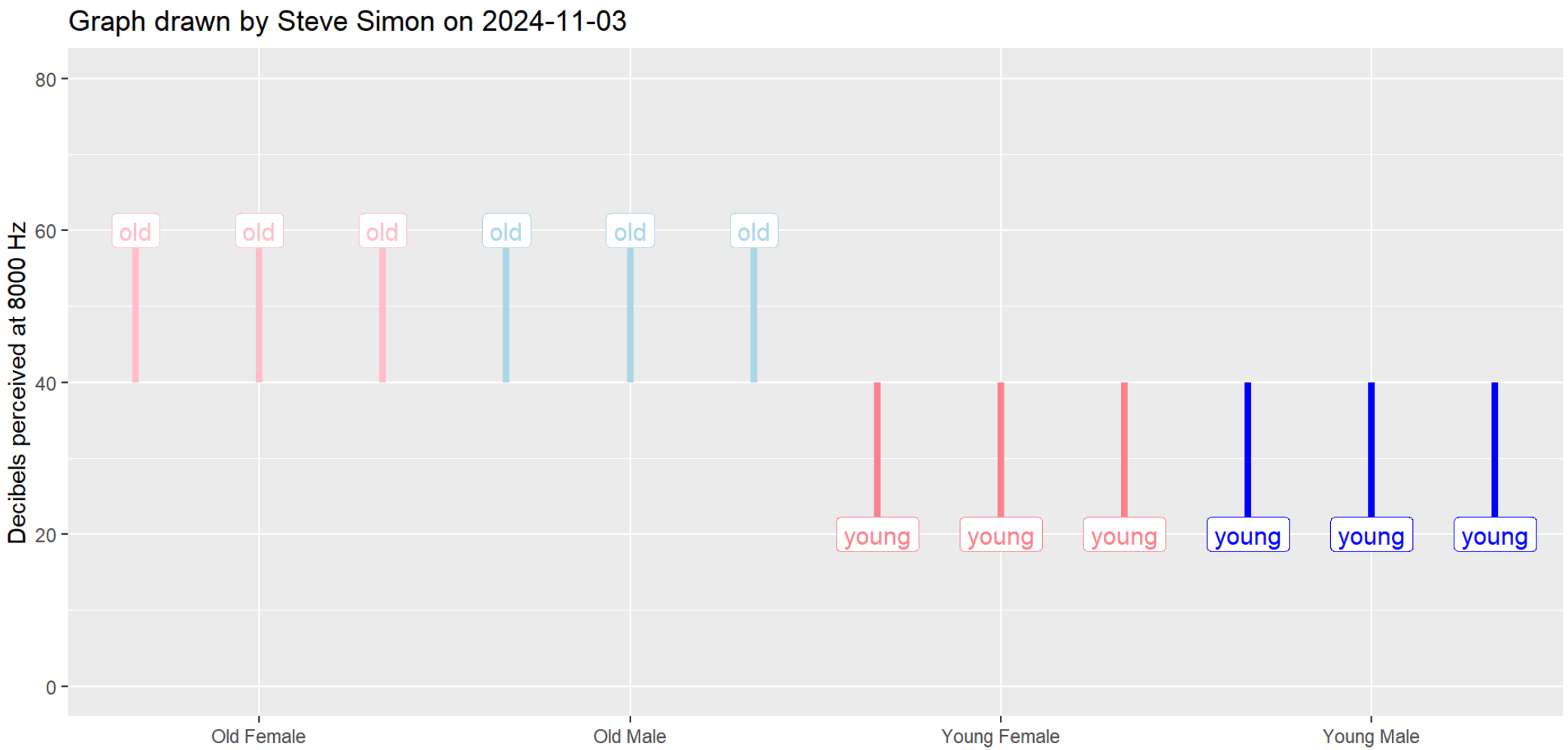
Artificial data with means

```
# A tibble: 12 × 8
      id age   gender code     db age_mean gender_mean overall_mean
   <int> <chr> <chr>  <chr> <dbl>    <dbl>       <dbl>        <dbl>
 1     1 old   female of       45       60          35           40
 2     2 old   female of       60       60          35           40
 3     3 old   female of       60       60          35           40
 4     4 old   male   om       65       60          45           40
 5     5 old   male   om       60       60          45           40
 6     6 old   male   om       70       60          45           40
 7     7 young female yf       20       20          35           40
 8     8 young female yf       20       20          35           40
 9     9 young female yf        5       20          35           40
10    10 young male   ym       25       20          45           40
11    11 young male   ym       20       20          45           40
```

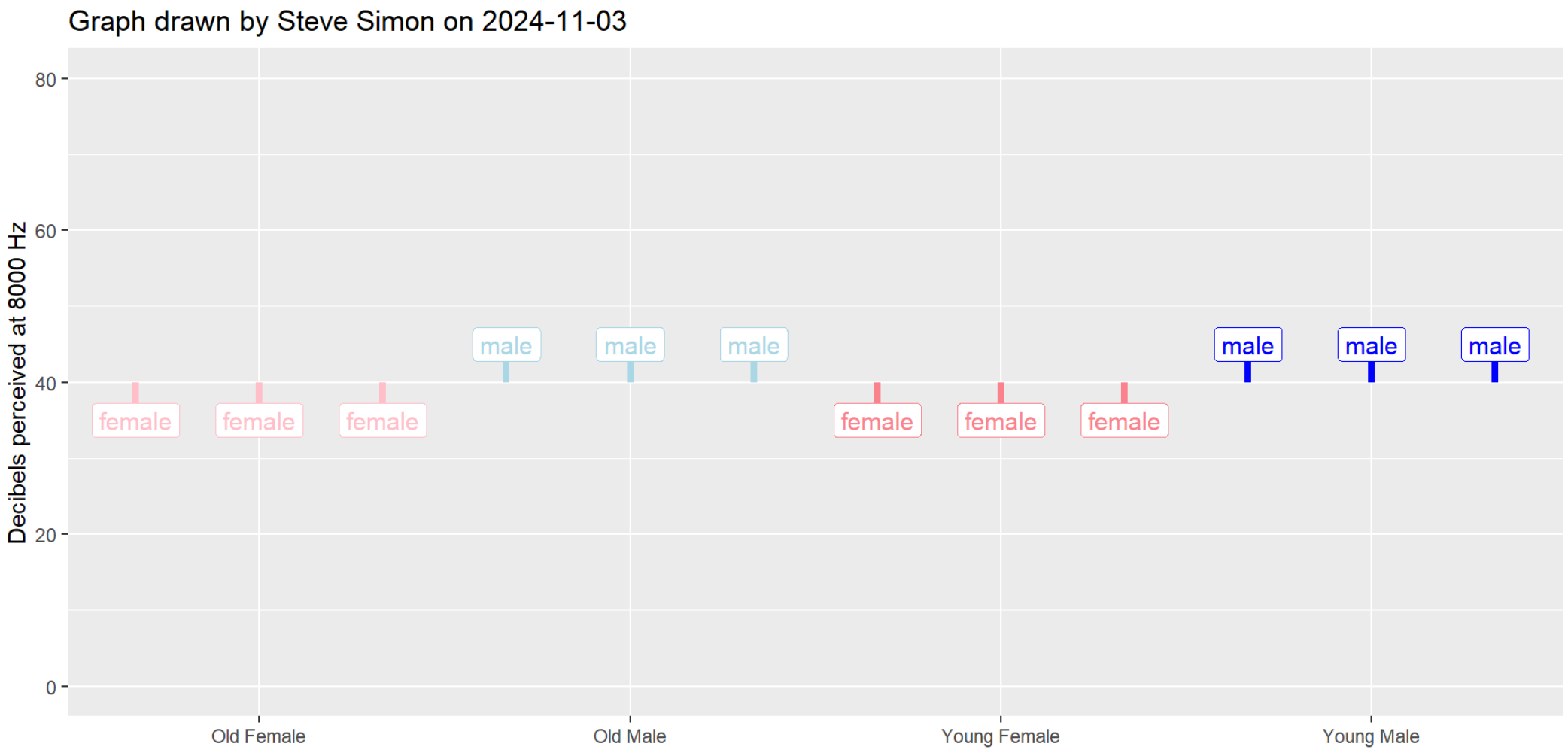
SS(Total)


Graph drawn by Steve Simon on 2024-11-03
Decibels perceived at 8000 Hz
80
60
40
20
0
of
of
of
om
om
om
yf
yf
yf
ym
ym
ym
Old Female
Old Male
Young Female
Young Male

SS(Age)


Graph drawn by Steve Simon on 2024-11-03
Decibels perceived at 8000 Hz
80
60
40
20
0
old
old
old
old
old
old
young
young
young
young
young
young
Old Female
Old Male
Young Female
Young Male

SS(Gender)


Graph drawn by Steve Simon on 2024-11-03
Decibels perceived at 8000 Hz
80
60
40
20
0
female
female
female
male
male
male
female
female
female
male
male
male
Old Female
Old Male
Young Female
Young Male

Analysis of variance table

```
Analysis of Variance Table

Response: db
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1   4800  4800.0  108.00 2.595e-06 ***
gender     1    300   300.0    6.75   0.02883 *
Residuals  9    400    44.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Break #1

- What you have learned
  - Two factor analysis of variance
- What’s coming next
  - Relationship to linear regression

Create indicator variables

```
# A tibble: 12 × 6
   age   gender code  i_young i_male    db
   <chr> <chr>  <chr>   <dbl>  <dbl> <dbl>
 1 old   female of          0      0    45
 2 old   female of          0      0    60
 3 old   female of          0      0    60
 4 old   male   om          0      1    65
 5 old   male   om          0      1    60
 6 old   male   om          0      1    70
 7 young female yf          1      0    20
 8 young female yf          1      0    20
 9 young female yf          1      0     5
10 young male   ym          1      1    25
11 young male   ym          1      1    20
```

Speaker notes

You need k-1 indicators for a categorical predictor that has k levels. In this simple example, that just means one indicator for age and one indicator for gender.

Two factor analysis of variance using aov

```r
m1 <- aov(db ~ age + gender, data=hearing)
anova(m1)
```

```
Analysis of Variance Table

Response: db
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1   4800  4800.0  108.00 2.595e-06 ***
gender     1    300   300.0    6.75   0.02883 *
Residuals  9    400    44.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Speaker notes

Here is a repeat of the analysis of variance table using aov.

Two factor analysis of variance using linear regression, 1

```r
m2 <- lm(db ~ age + gender, data=hearing)
anova(m2)
```

```
Analysis of Variance Table

Response: db
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1   4800  4800.0  108.00 2.595e-06 ***
gender     1    300   300.0    6.75   0.02883 *
Residuals  9    400    44.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Speaker notes

The analysis of variance table is identical when you use lm in place of aov.

Two factor analysis of variance using linear regression, 2

```
1 tidy(m2)
```

```
# A tibble: 3 × 5
  term        estimate std.error statistic      p.value
  <chr>          <dbl>     <dbl>     <dbl>        <dbl>
1 (Intercept)     55        3.33     16.5  0.0000000492
2 ageyoung       -40        3.85    -10.4  0.00000260
3 gendermale      10.0      3.85      2.60 0.0288
```

Speaker notes

The intercept is the estimated db when both indicator variables equal zero. The first indicator, i_age, is the estimated average change in db when moving from female to male. The second indicator, i_gender, is the estimated average change in db when moving from old to young.

Break #2

- What you have learned
  - Relationship to linear regression
- What’s coming next
  - Checking assumptions

Checking assumptions

- Normality (Non-normality)
- Homogeneity (Heterogeneity)
- Independence (Lack of independence)

Speaker notes

The assumptions for two factor analysis of variance is no different than for one factor analysis of variance or the two-sample t-test.

I’m a bit inconsistent in how I present this material. The assumptions are satisfied if you have normality, homogeneity, and independence. Equivalently, you could state that the assumptions are violated if you have non-normality or heterogeneity or lack of independence.

Use the boxplot to check assumptions

- Non-normality if boxplot shows skewness and/or outliers
- Heterogeneity if boxplot shows large change in variation
- Draw clustered boxplot to examine every combination of categories
  - Use a $\times$ b boxplots
- Independence is checked qualitatively

Alternatives if assumptions violated

- There is no analog to Kruskal-Wallis or Mann-Whitney-Wilcoxon
- Consider a log transformation
  - All values greater than 0
  - Groups with larger means have larger variation
  - Data is skewed right and outliers only on the high end

Break #3

- What you have learned
  - Checking assumptions
- What’s coming next
  - R code for two factor analysis of variance

Analysis of fruitfly data

Find the file simon-5501-12-moon.qmd on my github site.

Break #4

- What you have learned
  - R code for two factor analysis of variance
- What’s coming next
  - Interactions

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

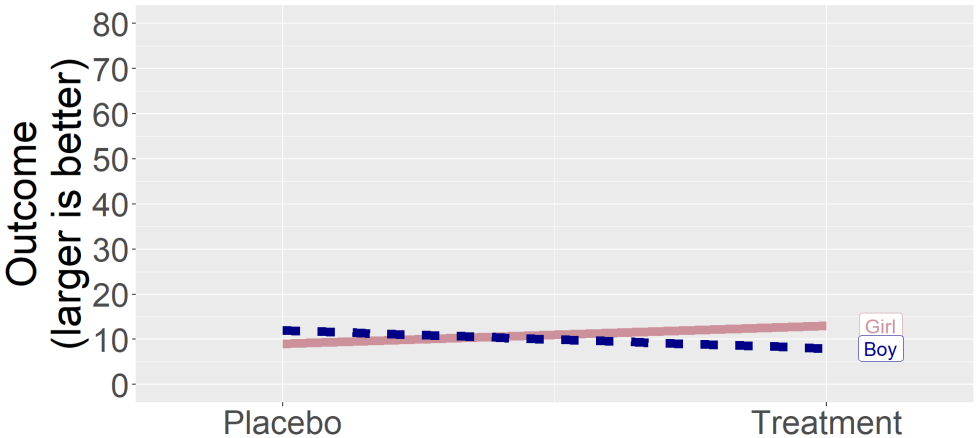
Speaker notes

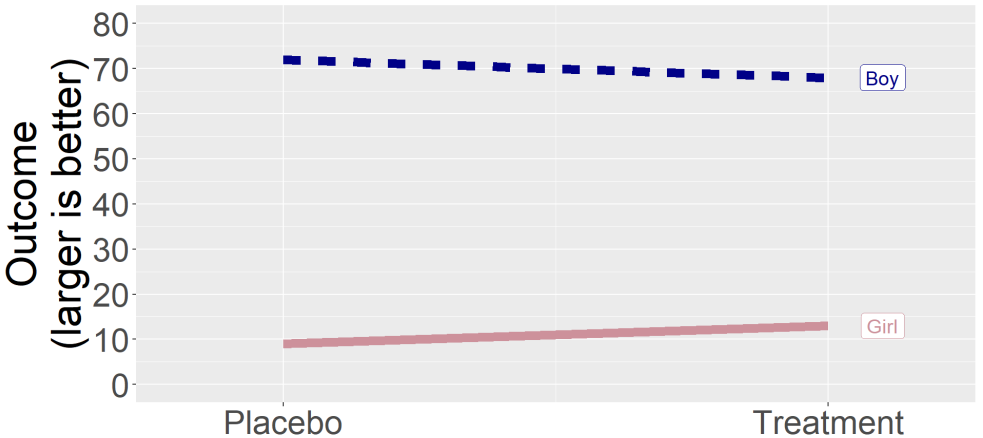
Interactions are important to look for, but if you find one, don’t rejoice. Interactions are a headache. They tell you that a simple interpretation of your research won’t work. That’s important to know, of course, but it also means that you will have to spend more time explaining your results in a paper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

Interaction plot

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

Hypothetical interaction plots, 1





- 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.

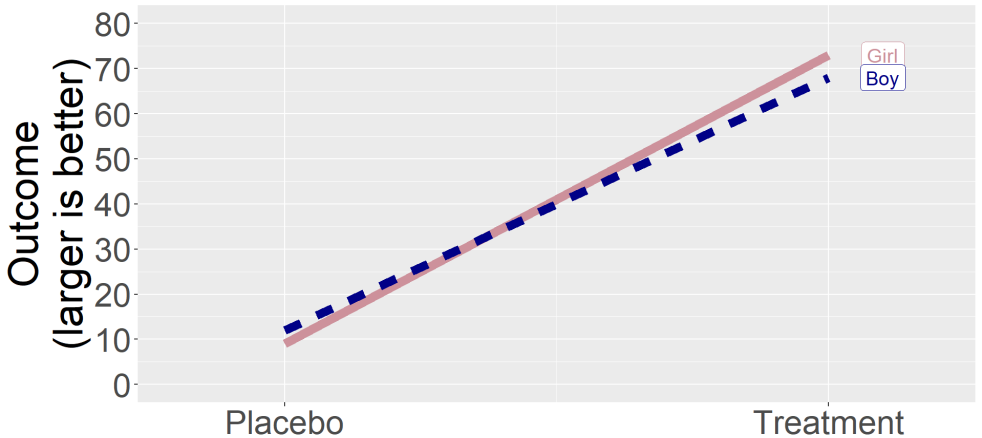
If the lines have roughly the same slope (both are flat or both are steep), then there is no interaction.

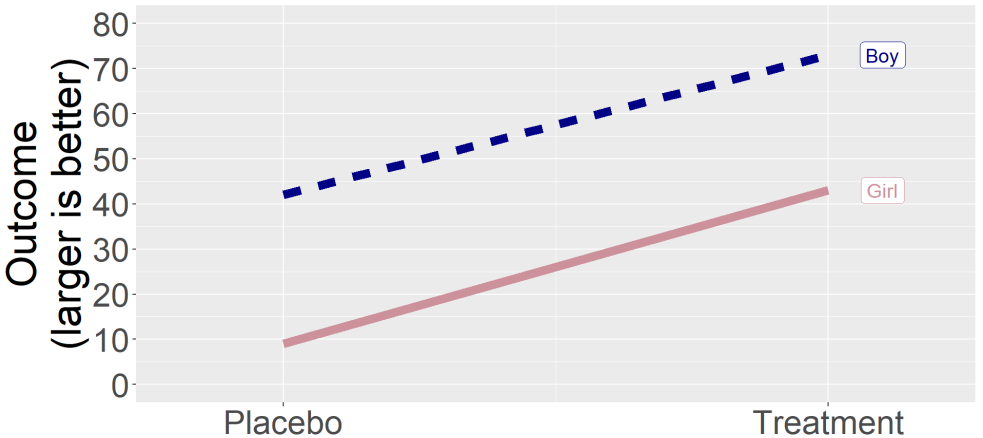
In the plot on the left, the two lines are flat, indicating that the treatment is ineffective. The outcome is not changed from the placebo.

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





- No interaction
- Effective treatment
- Boys/girls similar

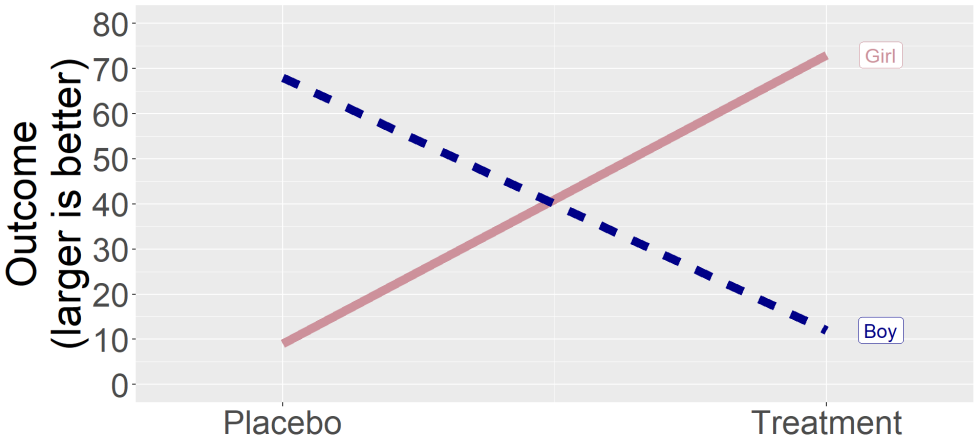
- No interaction
- Effective treatment
- Boys fare better than girls

Speaker notes

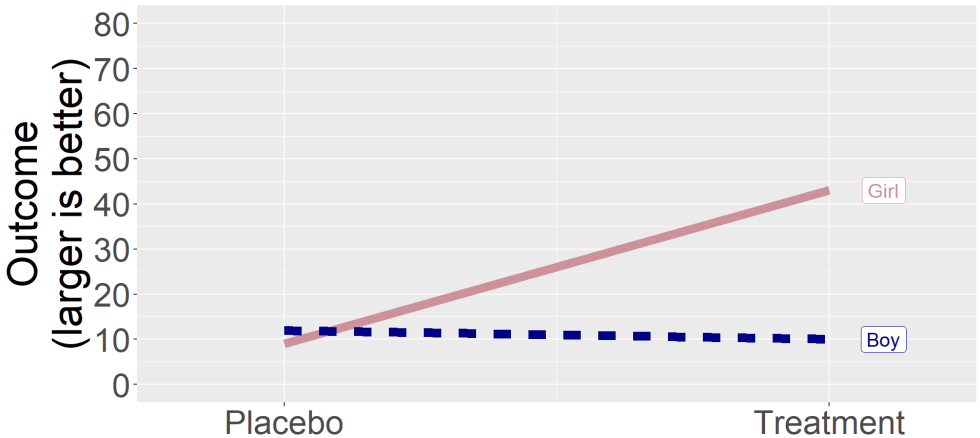
In the plot on the left, there is a steep slope for both boys and girls. The treatment is effective. There is no separation in the lines. Boys do not fare any better or worse on average than girls.

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



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



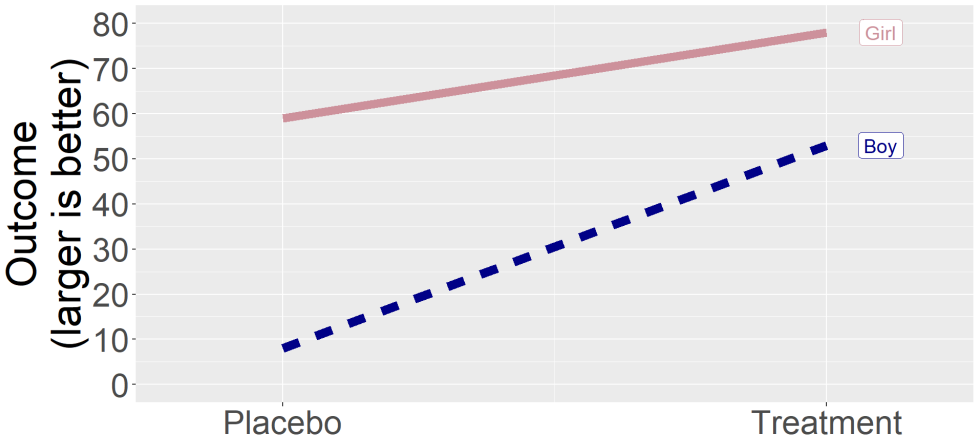
- Significant interaction
- Ineffective treatment in boys
- Effective treatment in girls

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



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

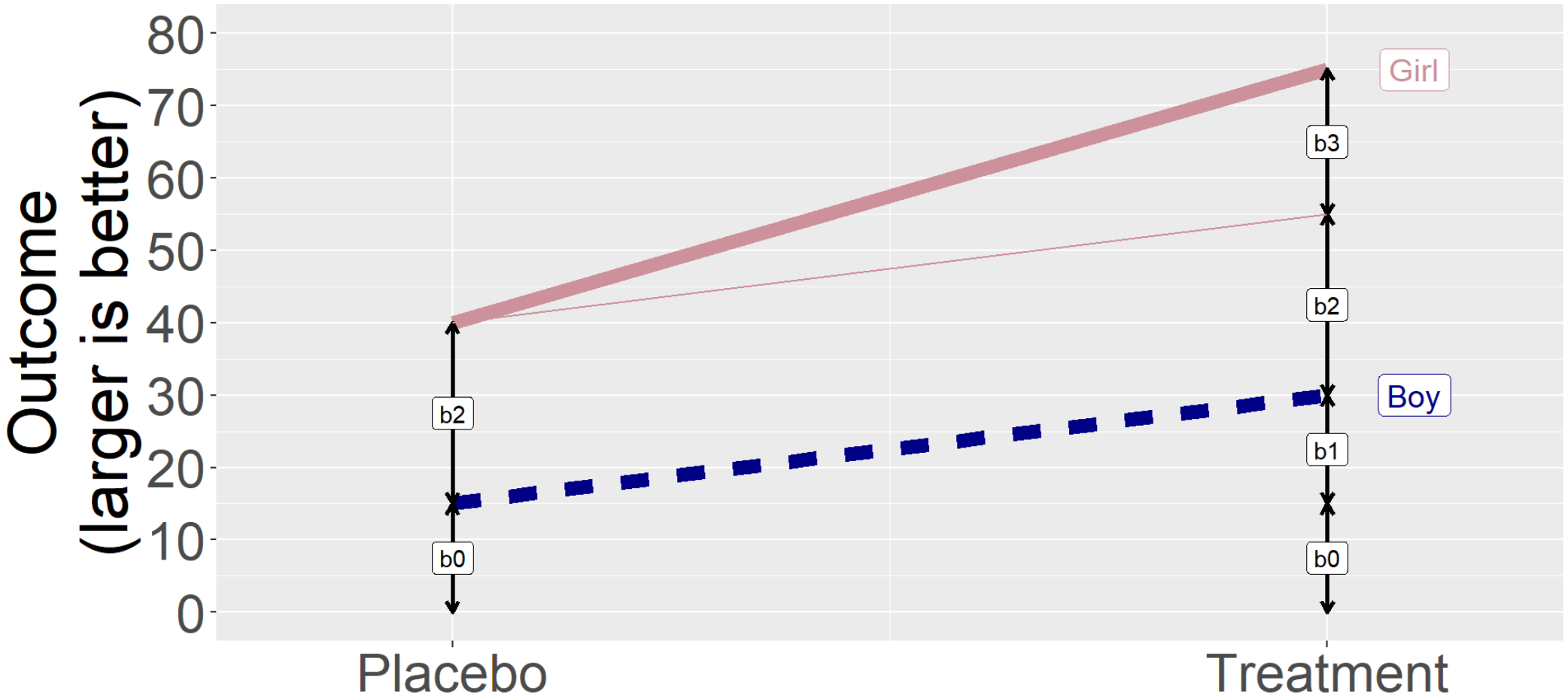
Speaker notes

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.

Indicator variable for an interaction

```
# A tibble: 12 × 7
   age   gender code  i_young i_male i_m_by_y    db
   <chr> <chr>  <chr>   <dbl>  <dbl>    <dbl> <dbl>
 1 old   female of          0      0        0    45
 2 old   female of          0      0        0    60
 3 old   female of          0      0        0    60
 4 old   male   om          0      1        0    65
 5 old   male   om          0      1        0    60
 6 old   male   om          0      1        0    70
 7 young female yf          1      0        0    20
 8 young female yf          1      0        0    20
 9 young female yf          1      0        0     5
10 young male   ym          1      1        1    25
11 young male   ym          1      1        1    20
```

Interpretation of intercept and slopes



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.

Example, full moon study, 1 of 2

```
# A tibble: 36 × 3
   month1 moon1      n
   <fct>  <fct>  <int>
 1 Aug    Before     1
 2 Aug    During     1
 3 Aug    After      1
 4 Sep    Before     1
 5 Sep    During     1
 6 Sep    After      1
 7 Oct    Before     1
 8 Oct    During     1
 9 Oct    After      1
10 Nov    Before     1
# ℹ 26 more rows
```

Speaker notes

Here is an example where you only have one observation for each combination of categories.

Example, full moon study, 2 of 2

```r
m1 <- aov(admission ~ month*moon, data=er)
anova(m1)
```

```
Analysis of Variance Table

Response: admission
           Df Sum Sq Mean Sq F value Pr(>F)
month      11 455.58  41.417     NaN    NaN
moon        2  41.51  20.757     NaN    NaN
month:moon 22 127.82   5.810     NaN    NaN
Residuals   0   0.00     NaN
```

Speaker notes

You lose two degrees of freedom for moon (3 phases: before, during, and after). You lose 11 degrees of freedom for month (12 months -1). The interaction has 2 times 11 or 22 degrees of freedom. You only started with 35 degrees of freedom. Subtract 2, 11, and 22, and you are left with zero degrees of freedom for error.

If you find yourself in this situation, just state that no test for interaction was possible in your methods section and highlight this as a weakness of your study in the discussion section.

Break #5

- What you have learned
  - Interactions
- What’s coming next
  - R code for interactions

Analysis of fruitfly data

Find the file simon-5501-12-fruitfly.qmd on my github site.

Break #6

- What you have learned
  - R code for interactions
- What’s coming next
  - Your homework

Your homework

Find the file simon-5501-12-directions.md on my github site.

Summary

- What you have learned
  - Two factor analysis of variance
  - Relationship to linear regression
  - Checking assumptions
  - R code for two factor analysis of variance
  - Interactions
  - R code for interactions
  - Your homework