MEDB 5501, Module12

2024-11-05

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.

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

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.

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.

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.

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, $\bar{Y}_{i..}$, the second factor, $\bar{Y}_{.j.}$, and an overall mean, $\bar{Y}_{...}$ which is an average across all of the data.

Mathematical model, 4

- $SS(Total) = \Sigma_i \Sigma_j \Sigma_k \left(Y_{ijk} - \bar{Y}_{...}\right)^2$
  - 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).

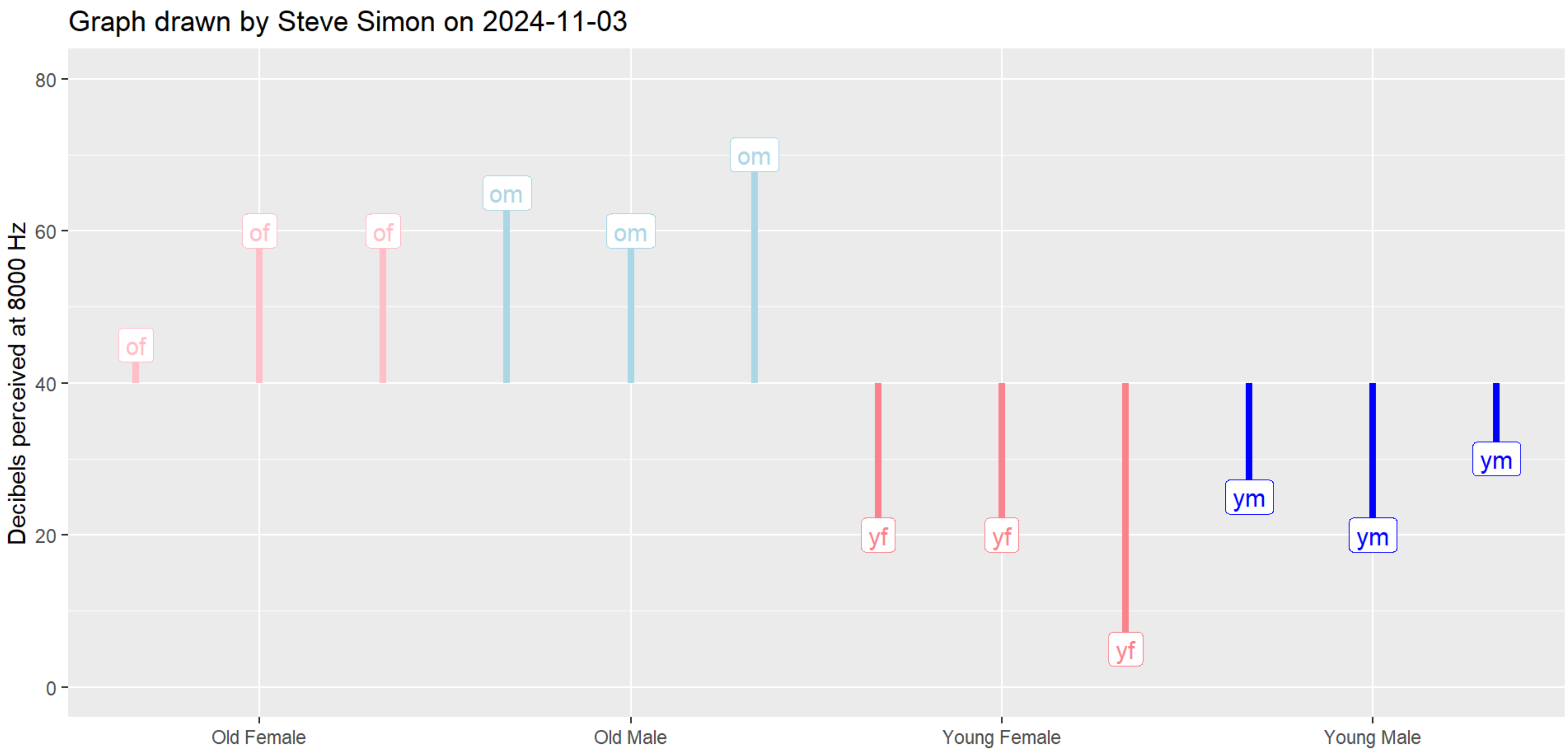
Artificial data

```
# 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
12      12 young male   ym       30
```

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
12    12 young male   ym       30       20          45           40
```

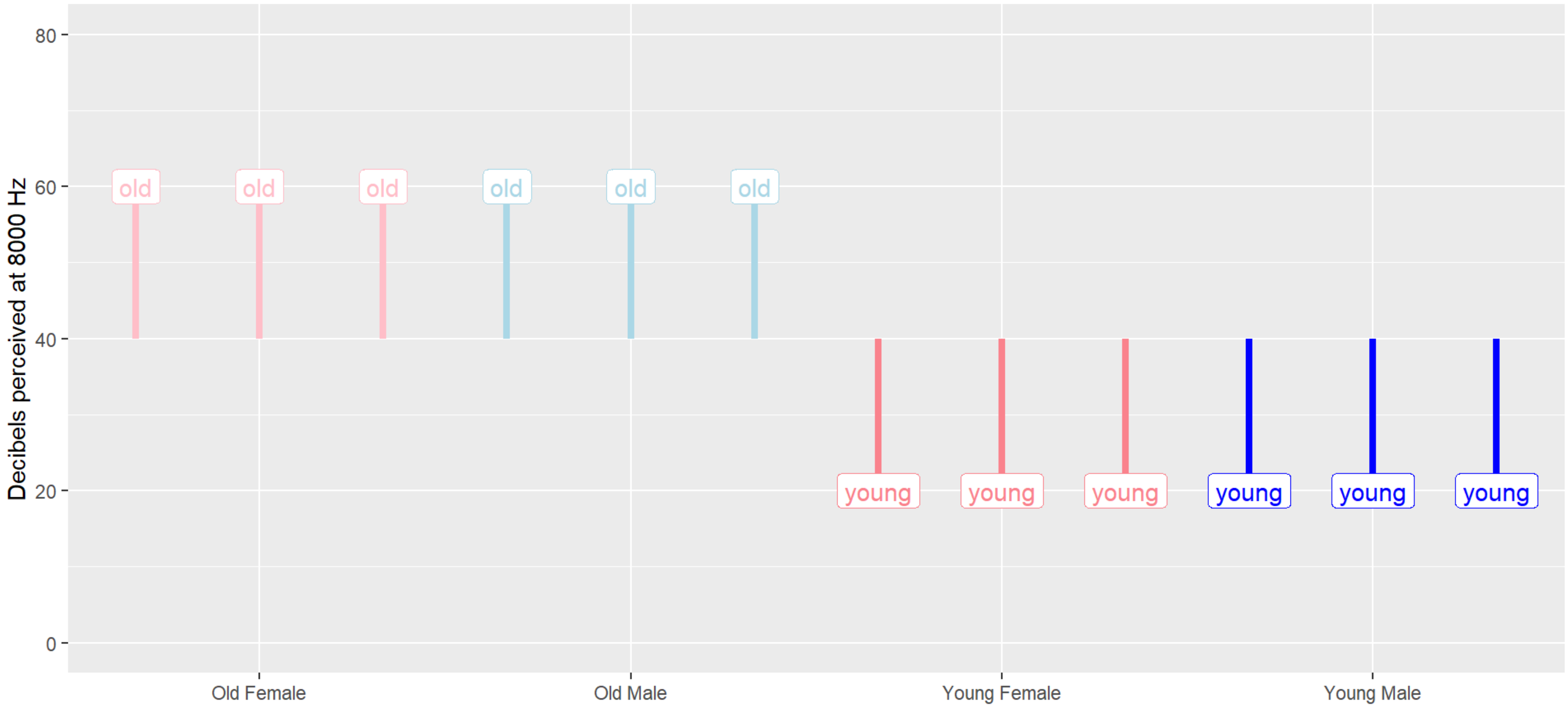
SS(Total)



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

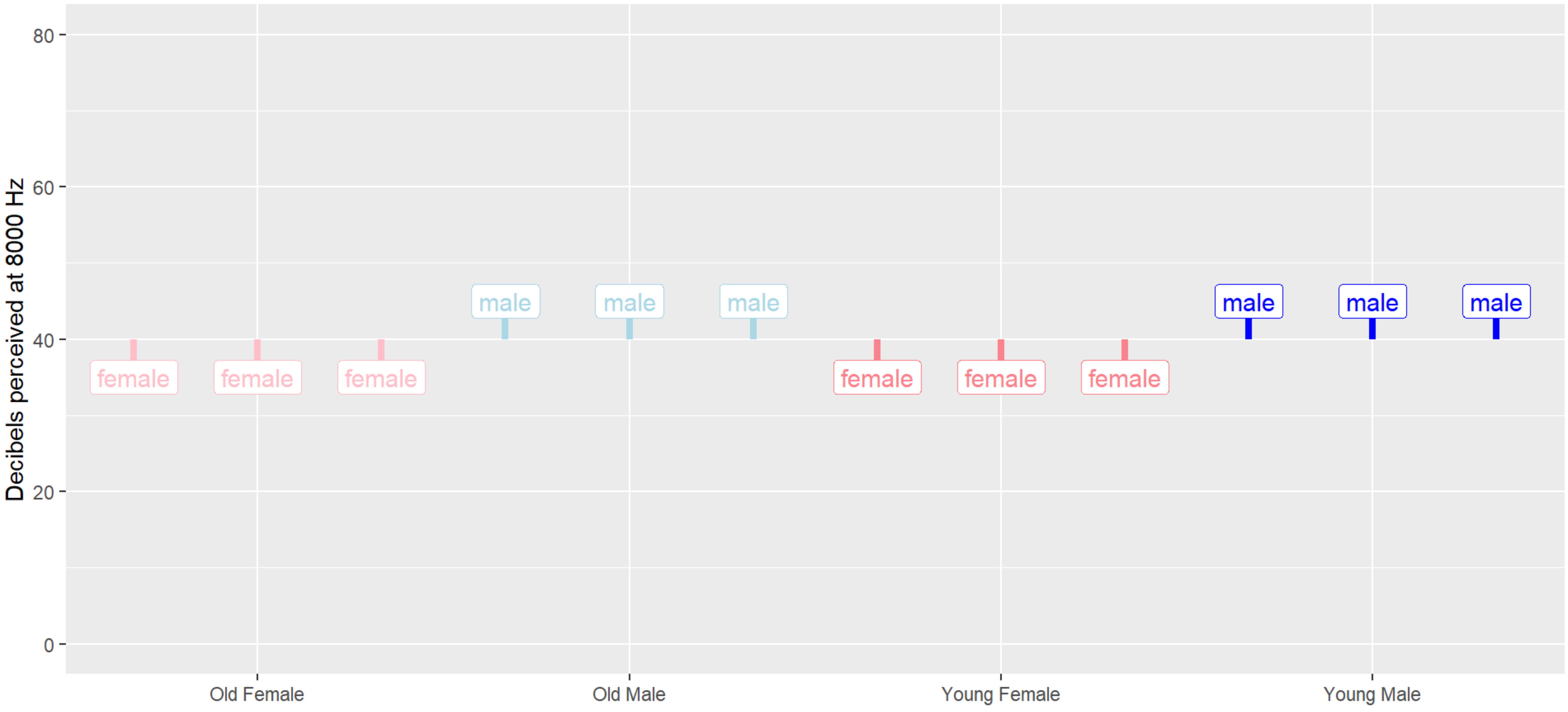
SS(Age)

Graph drawn by Steve Simon on 2024-11-03



SS(Gender)

Graph drawn by Steve Simon on 2024-11-03



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
12 young male   ym          1      1    30
```

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

```r
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

Listing of full-moon-er-admissions.yaml, 1

```
---
data_dictionary: full-moon.txt

description: >
  The data give the admission rates to the
  emergency room of a Virginia mental health
  clinic before, during and after the 12
  full moons from August 1971 to July 1972.

additional_description:
  https://gksmyth.github.io/ozdasl/general/fullmoon.html
```

Listing of full-moon-er-admissions.yaml, 2

```
source:
  - Blackman, S., and Catalina, D. (1973). The moon and the emergency room.
    Perceptual and Motor Skills 37, 624-626.
  - Ovlin, J. F. (1943). Moonlight and nervous disorders. American Journal of
    Psychiatry 99, 578-584.
  - Larsen, R.J., and Marx, M.L., (1986). An Introduction to Mathematical
    Statistics and Its Applications 2nd Edition. Prentice-Hall, Englewood
    Cliffs, New Jersey. Case Study 1.2.3.

copyright:
  unknown

format:
  tab-delimited
```

Listing of full-moon-er-admissions.yaml, 3

```
varnames:
  first row of data

size:
  rows: 36
  columns: 3

vars:
  Month:
    label: Month of year
    scale: Aug, Sep, ..., Jul
```

Listing of full-moon-er-admissions.yaml, 4

```
Moon:
  label: time relative to full moon
  values:
    - Before
    - During
    - After

Admission:
  label: Admission rate
  Unit: patients/day
```

Listing of simon-5501-12-moon.qmd, 1

```
---
title: "Analysis of full-moon-er-admissions data"
format:
  html:
    embed-resources: true
---

This program reads data on admissions to an emergency room before, during, and
after a full moon. Find more information in the [data dictionary][dd].

[dd]: https://github.com/pmean/data/blob/master/files/full-moon-er-
admissions.yaml

This code was written by Steve Simon on 2024-11-02 and is placed in the public
domain.
```

Listing of simon-5501-12-moon.qmd, 2

~~~
## Load the tidyverse library

```{r setup}
#| message: false
#| warning: false
library(broom)
library(tidyverse)
```

For most of your programs, you should load the tidyverse library. The broom
library converts your output to a nicely arranged dataframe. The messages and
warnings are suppressed.
~~~

Listing of simon-5501-12-moon.qmd, 3

~~~
## Read the data and view a brief summary

```{r read}
er <- read_tsv(
  "../data/full-moon-er-admissions.txt",
  col_types="ccn",
  col_names=TRUE)
names(er) <- tolower(names(er))
glimpse(er)
```

This is a tab delimited file. I changed all the variable names to all
lowercase.
That's one less thing opportunity for inconsistency errors.
~~~

Listing of simon-5501-12-moon.qmd, 4

~~~
## Reorder the category levels, 1

```{r reorder-moon}
er$moon1 <- factor(
  er$moon,
  levels=c(
    "Before",
    "During",
    "After"))
```

If you let R order the categories, it will do it alphabetically. The factor
function will place them in a more logical order.
~~~

Listing of simon-5501-12-moon.qmd, 5

~~~
## Reorder the category levels, 2

```{r reorder-month}
er$month1 <- factor(
  er$month,
  levels=c(
    "Aug", "Sep", "Oct",
    "Nov", "Dec", "Jan",
    "Feb", "Mar", "Apr",
    "May", "Jun", "Jul"))
```

Since the study started in August, it makes sense to order the months starting
then. An alphabetical ordering (April first then August?) makes no sense here.
~~~

Listing of simon-5501-12-moon.qmd, 6

~~~
## Draw boxplot of admission versus moon

```{r boxplots-admission-vs-moon}
#| fig.width: 6
#| fig.height: 2.0
er |>
  ggplot(aes(moon1, admission)) +
    geom_boxplot() +
    xlab("Moon phase") +
    ylab("ER admissions per day") +
    ggtitle("Graphs drawn by Steve Simon on 2024-11-02") +
    coord_flip()
```

Although there are a few outliers, I would not worry too much about the
normality assumption. It also looks like heterogeneity would not be too much
~~~

Listing of simon-5501-12-moon.qmd, 7

~~~
## Draw boxplot of admission versus month

```{r boxplots-admission-vs-month}
#| fig.width: 6
#| fig.height: 6.0
er |>
  ggplot(aes(month1, admission)) +
    geom_boxplot() +
    xlab("Month") +
    ylab("ER admissions per day") +
    ggtitle("Graphs drawn by Steve Simon on 2024-11-02") +
    coord_flip()
```

Although there are some unusual patterns here, keep in mind that there are
only two observations per month
~~~

Listing of simon-5501-12-moon.qmd, 8

~~~
## Calculate average ER admissions versus moon

```{r by-moon}
er |>
  group_by(moon1) |>
  summarize(
    admission_mn=mean(admission),
    admission_sd=sd(admission),
    n=n())
```

There is a slightly higher average admission rate and slightly more vaiation
during a full moon.
~~~

Listing of simon-5501-12-moon.qmd, 9

~~~
## Calculate average ER admissions versus month

```{r by-month}
er |>
  group_by(month1) |>
  summarize(
    admission_mn=mean(admission),
    admission_sd=sd(admission),
    n=n())
```

Er admissions are trending upward over time and there is some evidence of
heterogeneity. Again, treat this with caution as there are only three
observations per month.
~~~

Listing of simon-5501-12-moon.qmd, 10

~~~
## Analysis of variance

```{r anova}
m1 <- aov(admission ~ moon1 + month1, data=er)
anova(m1)
```

There is a borderline, but statistically significant difference in average
admission rates during different phases of the moon. There is a statistically
significant difference in average admission rates between some of the months.
~~~

Listing of simon-5501-12-moon.qmd, 11

~~~
## Using lm, 1

```{r lm-1}
m2 <- lm(admission ~ moon1 + month1, data=er)
anova(m2)
```

The analysis of variance table using lm is the same as the analysis of
variance
table using aov.
~~~

Listing of simon-5501-12-moon.qmd, 12

~~~
## Using lm, 2

```{r lm-2}
tidy(m2)
```

The intercept represents the estimated average admission rate before a full
moon
and during the month of August. The admission rate is 2.5 patients per day
higher during a full moon compared to a full moon. This is small but probably
still important. The difference in average admission rates is 0.54 patients
per
day higher after a full moon compared to before a full moon. The remaining
estimates compare each month back to the first month (August).
~~~

Listing of simon-5501-12-moon.qmd, 13

~~~
## Using lm, 3

```{r lm-3}
er |>
  group_by(moon1) |>
  summarize(admission_mn=mean(admission)) |>
  mutate(admission_before=first(admission_mn)) |>
  mutate(difference=admission_mn-admission_before)
```

This table shows how to calculate the regression coefficients associated with
phases of the moon. It is not really needed for the analysis, but helps
illustrate how these coefficients are computed.
~~~

Listing of simon-5501-12-moon.qmd, 14

~~~
## Using lm, 4

```{r lm-4}
er |>
  group_by(month1) |>
  summarize(admission_mn=mean(admission)) |>
  mutate(admission_aug=first(admission_mn)) |>
  mutate(difference=admission_mn-admission_aug)
```

This table shows how to calculate the regression coefficients associated with
months of the year.
~~~

Listing of simon-5501-12-moon.qmd, 15

~~~
## Save important files for later use

```{r save}
save(
  er,
  file="../data/full-moon-er-admissions.RData")
```
~~~

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

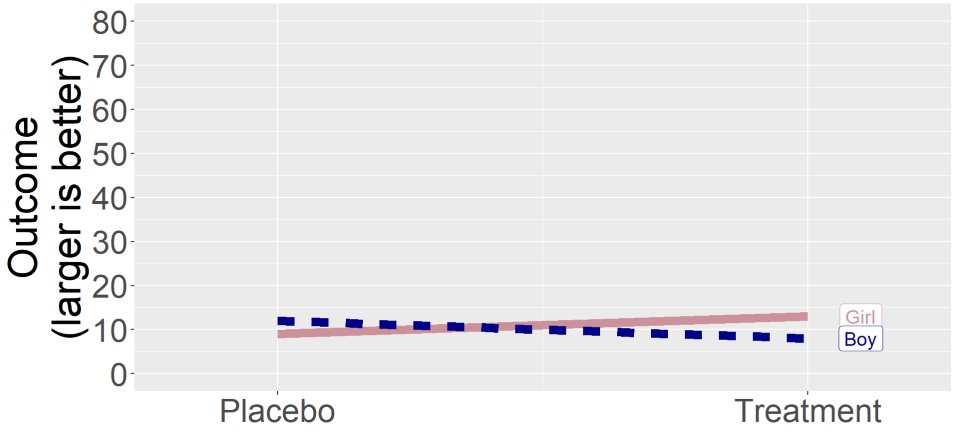
Speaker notes

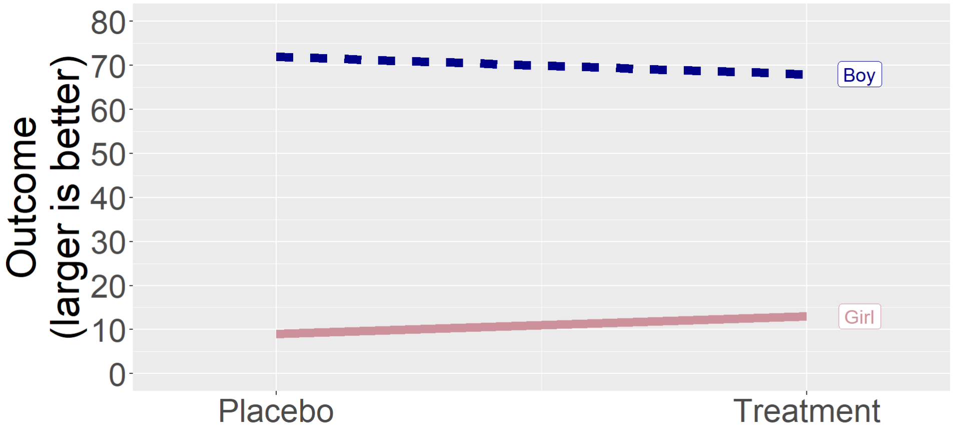
Interactions are important to look for, but if you find one, don’t rejoice. Interactions are a headache. They tell you that a simple interpretation of your research won’t work. That’s important to know, of course, but it also means that you will have to spend more time explaining your results in a paper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

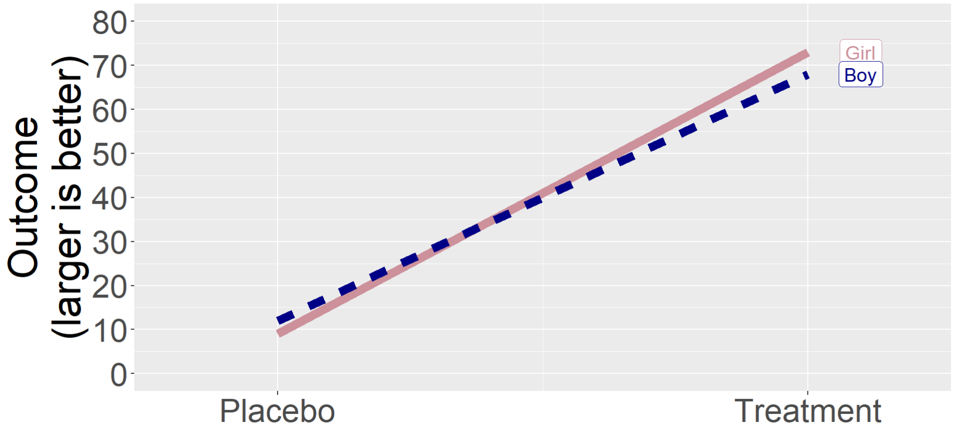
If the lines have roughly the same slope (both are flat or both are steep), then there is no interaction.

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.

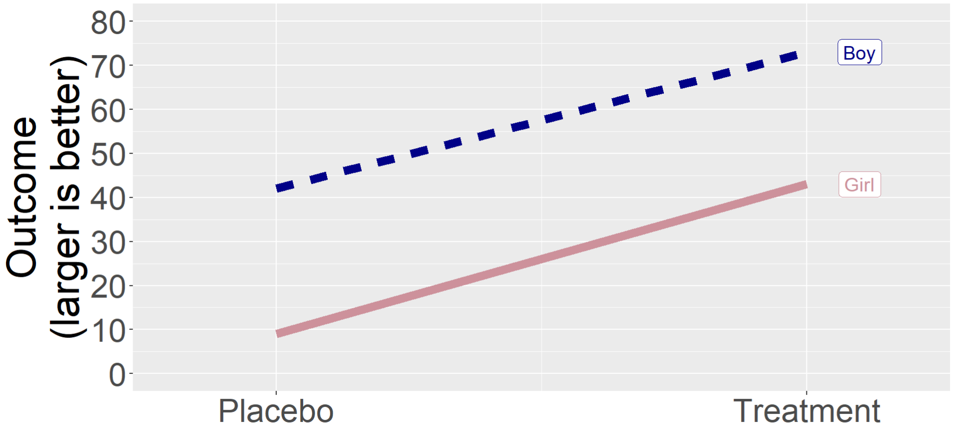
The two lines lie more or less on top of one another. This indicates that there is no difference in average outcome between boys are girls.

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



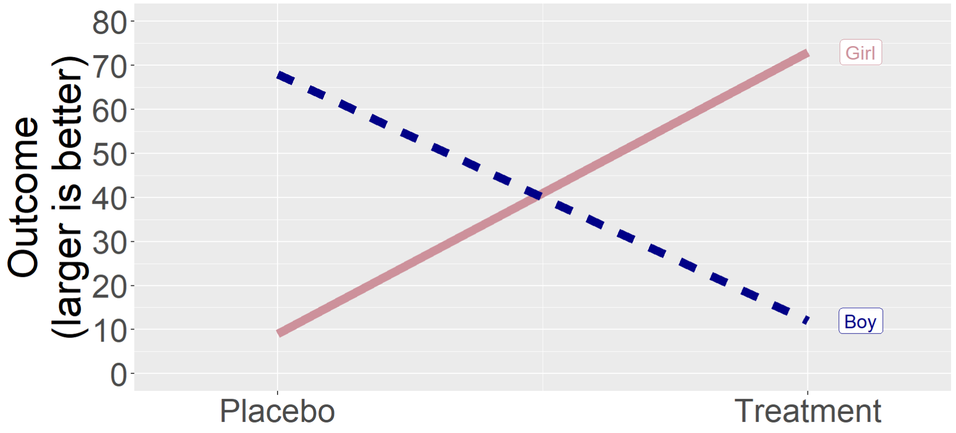
- No interaction
- Effective treatment
- Boys fare better than girls

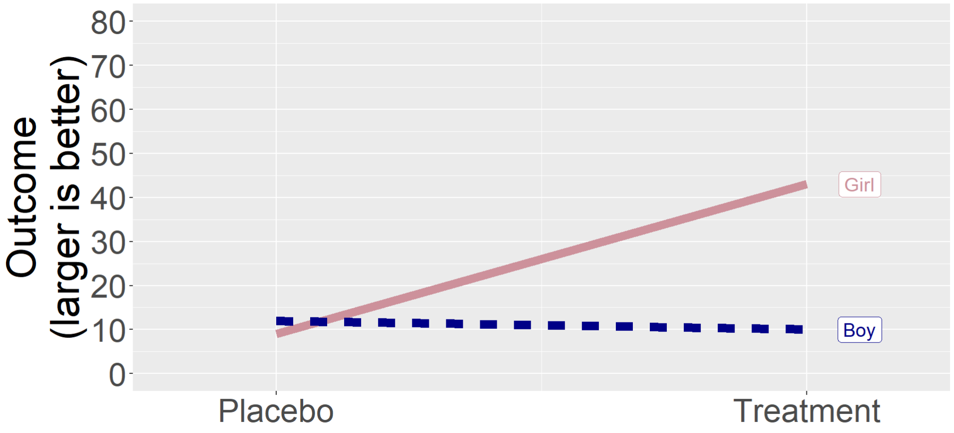
Speaker notes

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

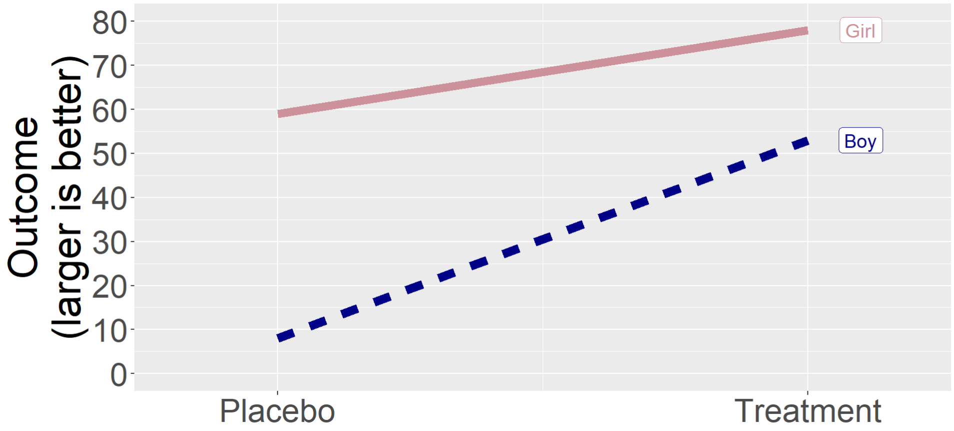
- Significant interaction
- Ineffective treatment in boys
- Effective treatment in girls

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

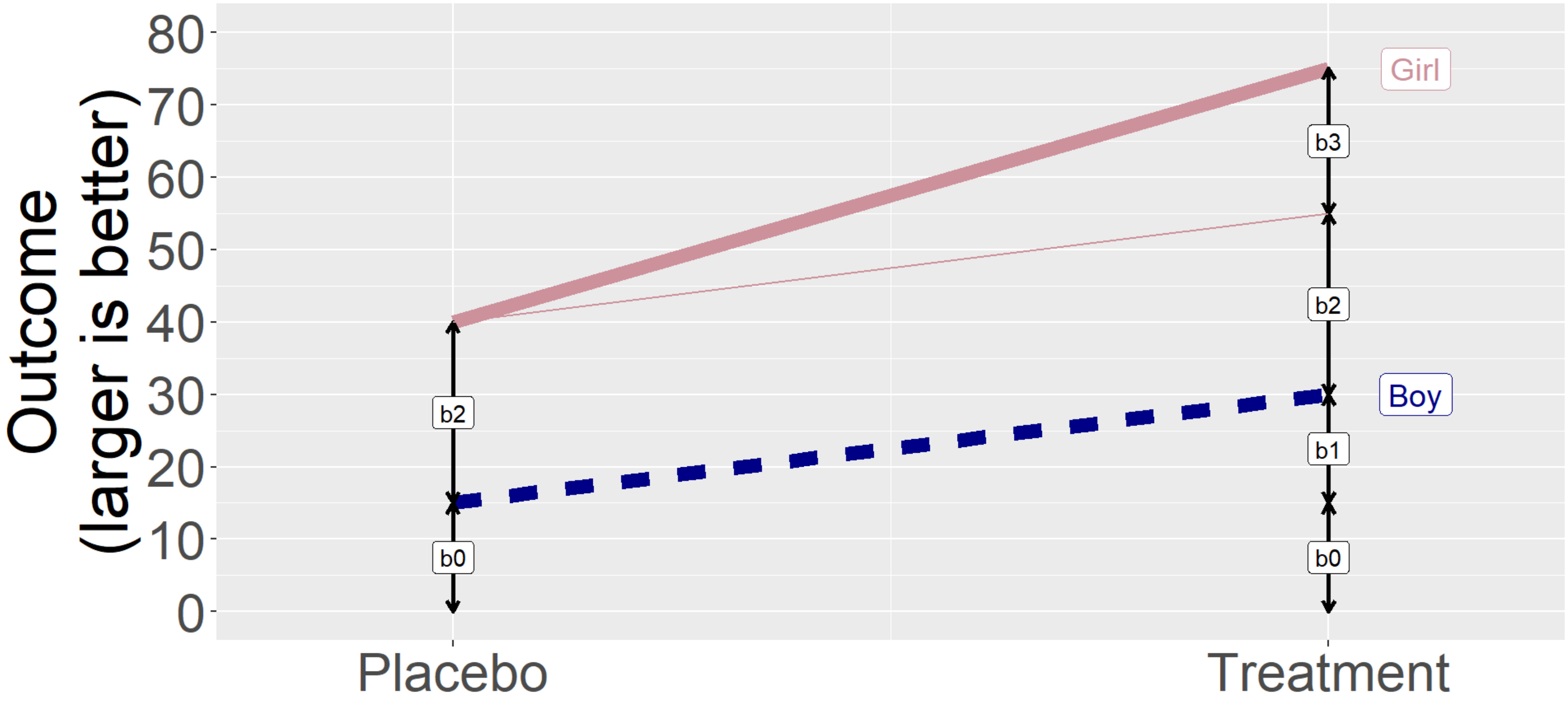
Speaker notes

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
12 young male   ym          1      1        1    30
```

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

Listing of fruitfly.yaml, 1

```
---
data_dictionary: fruitfly.dat.txt

copyright: >
  This dataset is copyrighted by the authors of the Journal of Statistics
  Education article, but should be available for individual educational uses
  under the Fair Use provisions of copyright law.

description: >
  Does access to mating affect the lifespan of fruitflies? This data shows the
  longevity of male fruitflies in the presence or absence of female fruitflies
  to mate with. Male fruitflies were housed with 0, 1, or 8 females. In some
  groups, the females were pregnant and thus not available for mating. There
are
  two covariates, length of the thorax and percentage of time sleeping, that
  might also influence longevity
```

Listing of fruitfly.yaml, 2

```
source:
  - Partridge, L., and Farquhar, M. (1981), "Sexual Activity and the Lifespan
of
    Male Fruitflies," Nature, 294, 580-581.
  - James A. Hanley & Stanley H. Shapiro. Sexual Activity and the Lifespan of
    Male Fruitflies: A Dataset That Gets Attention. Journal of Statistics
    Education v.2, n.1 (1994)

additional_description:
  https://jse.amstat.org/datasets/fruitfly.txt

download_url:
  https://jse.amstat.org/datasets/fruitfly.dat.txt
```

Listing of fruitfly.yaml, 3

```
format:
  fixed width

varnames:
  not included

missing_value_code:
  not needed

size:
  rows: 125
  columns: 6
```

Listing of fruitfly.yaml, 4

```
vars:
  id:
    columns: 1-2

  partners:
    columns: 4
    label: Number of female partners
    values: 0, 1, or 8
```

Listing of fruitfly.yaml, 5

```
type:
  columns: 6
  label: Type of female fruitfly
  values:
    0: newly pregnant female
    1: virgin female
    9: not applicable (when partners=0)

longevity:
  columns: 8-9
  label: Lifespan
  unit: days
  range: positive
```

Listing of fruitfly.yaml, 6

```
  thorax:
    columns: 11-14
    label: Length of thorax
    unit: mm
    range: positive

  sleep:
    columns: 16-17
    label: Percentage of each day sleeping
    range: 0 to 100
---
```

Listing of simon-5501-12-fruitfly.qmd, 1

```
---
title: "Analysis of fruitfly data"
format:
  html:
    embed-resources: true
---

This program reads data on fruit fly longevity. Find more information in the
[data dictionary][dd].

[dd]: https://github.com/pmean/data/blob/master/files/fruitfly.yaml

This code was written by Steve Simon on 2024-10-29 and is placed in the public
domain.
```

Listing of simon-5501-12-fruitfly.qmd, 2

~~~
## Load the tidyverse library

```{r setup}
#| message: false
#| warning: false
library(broom)
library(tidyverse)
```

#### Comments on the code

For most of your programs, you should load the tidyverse library. The broom
library converts your output to a nicely arranged dataframe. The messages and
warnings are suppressed.
~~~

Listing of simon-5501-12-fruitfly.qmd, 3

~~~
## List the variable names

```{r variable-list}
vlist <- c(
  "id",
  "partners",
  "type",
  "longevity",
  "thorax",
  "sleep")
```
~~~

Listing of simon-5501-12-fruitfly.qmd, 4

```
#### Comments on the code

When a dataset does not have variables on the first line, you need to specify
them in the code.
```

Listing of simon-5501-12-fruitfly.qmd, 5

~~~
## Read the data and view a brief summary

```{r read}
fly <- read_fwf(
  "../data/fruitfly.txt",
  col_types="nnnnnn",
  fwf_widths(
    widths=c(2, 2, 2, 3, 5, 3),
    col_names=vlist))
glimpse(fly)
```
~~~

Listing of simon-5501-12-fruitfly.qmd, 6

```
#### Comments on the code

The fruitfly dataset has a fixed width format (fwf). You need to specify the
columns that each variable uses.

Notice that the two categorical variables, partners and type, are actually
numbers rather than strings. To avoid having R treat these variables as if
they were continuous, use the factor function in some of the code below.
```

Listing of simon-5501-12-fruitfly.qmd, 7

~~~
## Create a subset of fruitfly dataset

```{r subset}
fly |>
  filter(type != 9) -> fly_subset
```
~~~

Listing of simon-5501-12-fruitfly.qmd, 8

```
#### Comments on the code

If you exclude the pure control group (No females), you can analyze the two
factors, partners and type individually. Partners has two category levels, 1
for
when one female was included in the cage and 8 for when eight females were
included in the cage. Type also has two category levels, 0 for pregnant female
fly/flies and 1 for virgin fly/flies. A male fly will not mate with a pregnant
females, so you can think of this as a second level of controls. The two
factors
are crossed, meaning that every possible combination of partners and type has
outcomes measured.
```

Listing of simon-5501-12-fruitfly.qmd, 9

~~~
## Draw boxplot of longevity by partners and type

```{r boxplots-partners-type}
#| fig.width: 6
#| fig.height: 2.5
fly_subset |>
  ggplot(aes(factor(partners), longevity, fill=factor(type))) +
    geom_boxplot() +
    xlab("Number of partners") +
    ylab("Lifespan in days") +
    ggtitle("Graphs drawn by Steve Simon on 2024-10-28") +
    labs(fill="Female fly type") +
    scale_fill_discrete(labels=c("Pregnant", "Virgin")) +
    coord_flip()
```
~~~

Listing of simon-5501-12-fruitfly.qmd, 10

```
#### Interpretation of the output

The lifespans tend to be quite a bit shorter in cages with 8 virgin females.
There are no problems with the homogeneity and normality assumptions.
```

Listing of simon-5501-12-fruitfly.qmd, 11

~~~
## Calculate average longevity by partners and type

```{r by-partners-and-type}
#| message: false
fly_subset |>
  group_by(type, partners) |>
  summarize(
    longevity_mn=mean(longevity),
    longevity_sd=sd(longevity),
    n=n()) -> fly_means
fly_means
```
~~~

Listing of simon-5501-12-fruitfly.qmd, 12

```
#### Interpretation of the output

The means show the same pattern as noted above. The standard deviations are
all
roughly the same.
```

Listing of simon-5501-12-fruitfly.qmd, 13

~~~
## Draw line graph

```{r line-graph}
fly_means |>
  ggplot(aes(
    factor(partners),
    longevity_mn,
    group=factor(type),
    color=factor(type))) +
    geom_line(linewidth=2) +
    expand_limits(y=range(fly_subset$longevity)) +
  xlab("Number of partners") +
  ylab("Lifespan in days") +
  labs(color="Female fly type") +
  scale_color_discrete(labels=c("Pregnant", "Virgin"))
```
~~~

Listing of simon-5501-12-fruitfly.qmd, 14

```
#### Interpretation of the output

This plot shows strong evidence of an interaction.

Male fruitflies do appear to have shorter average lifespans when sharing a
cage
with virgin females compared to pregnant females, but this effect is far
stronger when the number of female fruitflies is eight rather than one.

The line for pregnant (sharing a cage with pregnant females) is close to flat,
indicating that there is little impact of the number of pregnant females on
average longevity.
```

Listing of simon-5501-12-fruitfly.qmd, 15

~~~
## Analysis of variance including the interaction

```{r anova-with-interaction}
m1 <- aov(longevity ~ factor(partners)*factor(type), data=fly_subset)
anova(m1)
```

#### Interpretation of the output

There is a statistically significant interaction between the number of female
partners and the type of partners (pregnant or virgin). When you have an
interaction, you do not normally interpret the individual categorical
predictors.
~~~

Listing of simon-5501-12-fruitfly.qmd, 16

~~~
## Using lm, 1

```{r lm-with-interactions-1}
m2 <- lm(longevity ~ factor(partners)*factor(type), data=fly_subset)
anova(m2)
```

#### Interpretation of the output

The analysis of variance table is identical when using lm instead of aov.
~~~

Listing of simon-5501-12-fruitfly.qmd, 17

~~~
## Using lm, 2

```{r lm-with-interactions-2}
tidy(m2)
```

#### Interpretation of the output

The lifespan of fruitflies in cages with 8 virgin females is 17 days shorter
than what you would expect. If the effect of type and partners were
independent,
you'd expect to see a 9.4 (= 8.0 + 1.4) day decline on average. The
interaction
shows that this cage is actually 16.6 days worse than that on average.
~~~

Listing of simon-5501-12-fruitfly.qmd, 18

~~~
## Using lm, 3

```{r add-coefficients}
b <- coef(m2)
fly_means |>
  mutate(b0=b[1]) |>
  mutate(b1=b[2]) |>
  mutate(b2=b[3]) |>
  mutate(b3=b[4]) |>
    select(-longevity_sd, -n) -> fly_coefficients
fly_coefficients[1, 5:7] <- 0
fly_coefficients[2, 6:7] <- 0
fly_coefficients[3, c(5,7)] <- 0
fly_coefficients
```
~~~

Listing of simon-5501-12-fruitfly.qmd, 19

```
#### Comments on the code

The coef function takes the intercept and slopes from a linear regression
model
and stores them in a vector. The mutate command adds the individual intercept
and slopes to separate columns. Finally, place zeros in places where the
various
slopes are not needed.
```

Listing of simon-5501-12-fruitfly.qmd, 20

```
#### Interpretation of the output

This table is not needed for the data analysis, but is included to illustrate
how the regression coefficients are computed in a two factor analysis of
variance with an interaction.

The intercept is the estimated average with both categories each to the "zero"
or reference level. In this case, the average lifespan of flies in a cage with
one pregnant female.

The first slope coefficient (b1) represents the estimated average difference
in
lifespan between eight pregnant females and one pregnant females.

The second slope coefficient (b2) represents the estimated average difference
in
```

Listing of simon-5501-12-fruitfly.qmd, 21

~~~
## Save important files for later use

```{r save}
save(
  fly,
  fly_subset,
  file="../data/fruitfly.RData")
```

#### Comments on the code

It is always good form to save the important pieces of your data analysis that
might be re-used sometime later.
~~~

Break #6

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

Listing of simon-5501-12-directions.md, 1

```
---
title: "Directions for 5501-12 programming assignment"
---

This programming assignment was written by Steve Simon on 2024-10-08 and is
placed in the public domain.
```

Listing of simon-5501-12-directions.md, 2

```
## Program

-   Download the [program][tem]
    -   Store it in your src folder
-   Modify the file name
    -   Use your last name instead of "simon"
-   Modify the documentation header
    -   Add your name to the author field
    -   Optional: change the copyright statement

[tem]: https://github.com/pmean/classes/blob/master/general/simon-5501-12-fruitfly.md
```

Listing of simon-5501-12-directions.md, 3

```
## Data

-   Download the [data][dat] file
    -   Store it in your data folder
-   Refer to the [data dictionary][dic], if needed.

[dat]: https://github.com/pmean/data/blob/main/files/fruitfly.txt
[dic]: https://github.com/pmean/data/blob/main/files/fruitfly.yaml
```

Listing of simon-5501-12-directions.md, 4

```
## Question 1

Create a subset of the fruitfly data by removing the age where type equals 9.
Draw a clustered boxplot with sleep as the outcome and partners and type as
the
categorical predictors. Interpret this graph. Is there evidence of
non-normality?

## Question 2

Calculate descriptive statistics for sleep (mean, standard deviation, and
sample size) by the combination of the two categorical predictors, partners
and
type. Is there evidence of heterogeneity?
```

Listing of simon-5501-12-directions.md, 5

```
## Question 3

Draw a line graph for the mean sleep levels compared by type and partners. Is
there evidence of an interaction?

## Question 4

Analyze the sleep variable using a two factor analysis of variance with an
interaction. Present and interpret the analysis of variance table.
```

Listing of simon-5501-12-directions.md, 6

```
## Question 5

What factors might make you consider using a log transformation for the sleep
variable? Do not run such an analysis but tell us whether you think the data
would warrant such a transformation?
```

Listing of simon-5501-12-directions.md, 7

```
## Your submission

-   Save the output in html format
-   Convert it to pdf format.
-   Make sure that the pdf file includes
    -   Your last name
    -   The number of this course
    -   The number of this module
-   Upload the file
-   Please note the [policy on late submissions and rework][sim3].

[sim3]: https://github.com/pmean/classes/blob/master/general/policy-on-extensions-and-rework.md
```

Listing of simon-5501-12-directions.md, 8

```
## If it doesn't work

Please review the [suggestions if you encounter an error page][sim4].

[sim4]: https://github.com/pmean/classes/blob/master/general/suggestions-if-you-encounter-an-error.md
```

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