431 Class 18

https://thomaselove.github.io/431-2024/

2024-10-29

## Today’s Agenda

Two-Factor (Two-Way) Analysis of Variance

* The importance of interaction
* Interpreting ANOVA output
* What are the key model checks for ANOVA?
* Comparing Model Performance Indices
* What if we add a covariate?

## R Packages

library(janitor)

Attaching package: 'janitor'

The following objects are masked from 'package:stats':  
  
 chisq.test, fisher.test

library(readxl)  
library(easystats)

# Attaching packages: easystats 0.7.3  
✔ bayestestR 0.15.0 ✔ correlation 0.8.6   
✔ datawizard 0.13.0 ✔ effectsize 0.8.9   
✔ insight 0.20.5 ✔ modelbased 0.8.9   
✔ performance 0.12.4 ✔ parameters 0.23.0  
✔ report 0.5.9 ✔ see 0.9.0

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.5.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
✔ purrr 1.0.2

── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

source("c18/data/Love-431.R")  
knitr::opts\_chunk$set(comment = NA)  
  
theme\_set(theme\_bw())

## The dm464 data, in full

We’ve discussed subsets of these data in previous classes.

dm464 <- read\_xlsx("c18/data/dm464.xlsx") |>  
 janitor::clean\_names() |>  
 mutate(across(where(is.character), as\_factor)) |>  
 mutate(id\_code = as.character(id\_code),  
 smoke\_start = fct\_relevel(smoke\_start, "Current", "Former"),  
 smoke\_end = fct\_relevel(smoke\_end, "Current", "Former"),  
 insur\_start = fct\_relevel(insur\_start, "Medicare",   
 "Commercial", "Medicaid"),  
 insur\_end = fct\_relevel(insur\_end, "Medicare",   
 "Commercial", "Medicaid"))  
  
dim(dm464)

[1] 464 42

## Descriptions for Today’s Variables

* Cohort of 464 English-speaking adults with diabetes measured at the start and end of a two-year period.

| Variable | Description |
| --- | --- |
| id\_code | DM-xxxx code to identify subjects (not consecutive) |
| bmi\_end | Body Mass Index in at the end of the period |
| dep\_start | Depression diagnosis at the start of the period (1 = yes, 0 = no) |
| ed\_start | High = more than 78% of adult residents of subject’s home neighborhood graduated high school, vs. Low = no more than 78%. |
| insur\_start | Subject’s insurance status at the start (3 levels) |

* 78% was close to the median value of the quantitative measure of high school graduation rate among enrolled subjects at the start of the period.
* All subjects have either Medicare, Commercial or Medicaid insurance.

## Today’s Variables

dm18 <- dm464 |>  
 mutate(ed\_start = factor(ifelse(nedlev\_start > 78, "High", "Low")),  
 dep\_start = factor(depdiag\_start)) |>  
 select(id\_code, bmi\_end, dep\_start, ed\_start, insur\_start)  
  
head(dm18)

# A tibble: 6 × 5  
 id\_code bmi\_end dep\_start ed\_start insur\_start  
 <chr> <dbl> <fct> <fct> <fct>   
1 DM-1007 47.2 0 Low Medicare   
2 DM-1017 34.7 0 High Commercial   
3 DM-1039 50 1 High Medicare   
4 DM-1049 57.2 1 Low Medicaid   
5 DM-1064 45.2 1 Low Medicaid   
6 DM-1070 34.3 1 High Medicaid

## data\_codebook() results

data\_codebook(dm18 |> select (-id\_code))

select(dm18, -id\_code) (464 rows and 4 variables, 4 shown)  
  
ID | Name | Type | Missings | Values | N  
---+-------------+-------------+----------+------------+------------  
1 | bmi\_end | numeric | 0 (0.0%) | [16, 78.3] | 464  
---+-------------+-------------+----------+------------+------------  
2 | dep\_start | categorical | 0 (0.0%) | 0 | 295 (63.6%)  
 | | | | 1 | 169 (36.4%)  
---+-------------+-------------+----------+------------+------------  
3 | ed\_start | categorical | 0 (0.0%) | High | 228 (49.1%)  
 | | | | Low | 236 (50.9%)  
---+-------------+-------------+----------+------------+------------  
4 | insur\_start | categorical | 0 (0.0%) | Medicare | 193 (41.6%)  
 | | | | Commercial | 61 (13.1%)  
 | | | | Medicaid | 210 (45.3%)  
--------------------------------------------------------------------

# Two-Factor Analysis of Variance: Example 1

## Example 1

Suppose we want to simultaneously understand the impacts of two factors on BMI at the end of the period, specifically:

* dep\_start (1 or 0), and
* ed\_start (Low or High)

and it is possible that the impact of depression diagnosis on BMI may depend on the home neighborhood’s educational attainment, and vice versa (i.e. depression and education may **interact**.)

## An interaction plot is a plot of means.

Calculate and store the group means.

ex1\_means <- dm18 |>  
 group\_by(dep\_start, ed\_start) |>  
 summarize(mean\_bmi = mean(bmi\_end))

`summarise()` has grouped output by 'dep\_start'. You can override using the  
`.groups` argument.

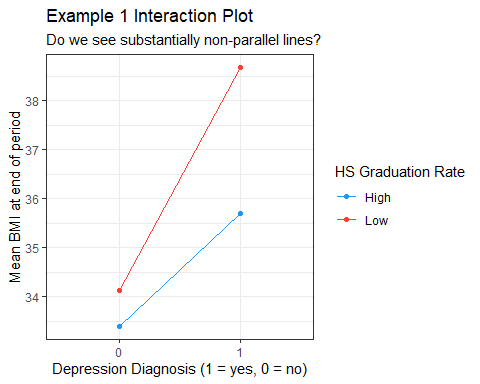
ex1\_means

# A tibble: 4 × 3  
# Groups: dep\_start [2]  
 dep\_start ed\_start mean\_bmi  
 <fct> <fct> <dbl>  
1 0 High 33.4  
2 0 Low 34.1  
3 1 High 35.7  
4 1 Low 38.7

## How strong is the interaction?

Now, we’ll plot these means, and look for a substantial interaction (non-parallel lines.)

ggplot(ex1\_means, aes(x = dep\_start, y = mean\_bmi)) +  
 geom\_line(aes(group = ed\_start, color = ed\_start)) +  
 geom\_point(aes(color = ed\_start)) +  
 scale\_color\_material() +  
 labs(title = "Example 1 Interaction Plot",  
 subtitle = "Do we see substantially non-parallel lines?",  
 y = "Mean BMI at end of period",   
 x = "Depression Diagnosis (1 = yes, 0 = no)",   
 color = "HS Graduation Rate")



## Two-Way ANOVA without interaction

This model predicts our outcome (bmi\_end) using the two factors as main effects only (with no interaction.)

mainfx <- lm(bmi\_end ~ dep\_start + ed\_start, data = dm18)  
  
mainfx

Call:  
lm(formula = bmi\_end ~ dep\_start + ed\_start, data = dm18)  
  
Coefficients:  
(Intercept) dep\_start1 ed\_startLow   
 32.981 3.452 1.555

## Interpret the estimate: 3.45?

model\_parameters(mainfx, ci = 0.90)

Parameter | Coefficient | SE | 90% CI | t(461) | p  
----------------------------------------------------------------------  
(Intercept) | 32.98 | 0.65 | [31.91, 34.05] | 50.68 | < .001  
dep start [1] | 3.45 | 0.84 | [ 2.06, 4.84] | 4.10 | < .001  
ed start [Low] | 1.56 | 0.81 | [ 0.22, 2.89] | 1.92 | 0.056

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

* If we have two subjects who live in neighborhoods with the same ed\_start value, but only one has a depression diagnosis at the start of the period, then **on average** the subject with a depression diagnosis will have a BMI that is 3.45 higher than the subject without such a diagnosis, according to our mainfx model.

## Interpret the estimate: 1.56?

model\_parameters(mainfx, ci = 0.90)

Parameter | Coefficient | SE | 90% CI | t(461) | p  
----------------------------------------------------------------------  
(Intercept) | 32.98 | 0.65 | [31.91, 34.05] | 50.68 | < .001  
dep start [1] | 3.45 | 0.84 | [ 2.06, 4.84] | 4.10 | < .001  
ed start [Low] | 1.56 | 0.81 | [ 0.22, 2.89] | 1.92 | 0.056

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

* Our mainfx model estimates that if we have two subjects who have the same depression diagnosis status at the start of the period, but live in neighborhoods with different ed\_start values (one High and one Low), then **on average** the subject living in the Low education neighborhood will have a BMI that is 1.56 higher than the other subject.

## Interpret the intercept term?

model\_parameters(mainfx, ci = 0.90)

Parameter | Coefficient | SE | 90% CI | t(461) | p  
----------------------------------------------------------------------  
(Intercept) | 32.98 | 0.65 | [31.91, 34.05] | 50.68 | < .001  
dep start [1] | 3.45 | 0.84 | [ 2.06, 4.84] | 4.10 | < .001  
ed start [Low] | 1.56 | 0.81 | [ 0.22, 2.89] | 1.92 | 0.056

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

* 32.98 is the intercept. It’s the mainfx model’s estimate of the mean of bmi\_end across all subjects where both dep\_start is 0 and ed\_start = High.

## ANOVA table for main effects model

anova(mainfx)

Analysis of Variance Table  
  
Response: bmi\_end  
 Df Sum Sq Mean Sq F value Pr(>F)   
dep\_start 1 1303 1302.59 17.1020 4.21e-05 \*\*\*  
ed\_start 1 280 280.48 3.6824 0.05561 .   
Residuals 461 35113 76.17   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### % of Total SS explained by the two factors?

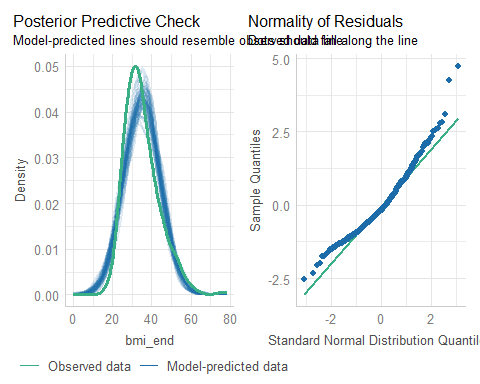
* Total SS = 1303 + 280 + 35113 = 36696
* So % explained = (1303 + 280) / 36696 = 0.043, or 4.3%

## Model mainfx and its assumptions

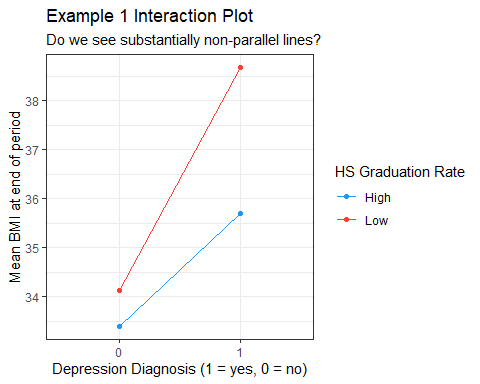
* mainfx is a “main effects only” model. It assumes there is zero interaction between dep\_start and ed\_start in predicting our outcome, bmi\_end.
* Like any ANOVA model, mainfx makes the usual linear model assumptions (linearity, constant variance, Normality)
  + Focus: posterior predictive check, and Normal Q-Q plot of residuals[[1]](#footnote-38).

## Checking model mainfx

set.seed(20241029)  
check\_model(mainfx, check = c("pp\_check", "qq"), detrend = FALSE)



## Interaction Plot, again



## Two-Way ANOVA with Interaction

Fit the linear model, including an interaction (\*) between our factors.

withint <- lm(bmi\_end ~ dep\_start \* ed\_start, data = dm18)  
  
withint

Call:  
lm(formula = bmi\_end ~ dep\_start \* ed\_start, data = dm18)  
  
Coefficients:  
 (Intercept) dep\_start1 ed\_startLow   
 33.3878 2.3073 0.7447   
dep\_start1:ed\_startLow   
 2.2284

## What is the withint model equation?

$$
bmi\_{end} = 33.4 + 2.3 (dep\_{start} = 1) \\
+ 0.7 (ed\_{start} = Low) \\ + 2.2 (dep\_{start} = 1) \times (ed\_{start} = Low)
$$

| dep\_start | ed\_start | Estimated bmi\_end |
| --- | --- | --- |
| 0 | High | 33.4 |
| 0 | Low | 33.4 + 0.7 = 34.1 |
| 1 | High | 33.4 + 2.3 = 35.7 |
| 1 | Low | 33.4 + 2.3 + 0.7 + 2.2 = 38.6 |

## Effect of High vs. Low ed\_start?

| Est. bmi\_end | ed\_start = High | ed\_start = Low |
| --- | --- | --- |
| dep\_start = 0 | 33.4 | 34.1 |
| dep\_start = 1 | 35.7 | 38.6 |

* If dep\_start = 0, High ed\_start yields 33.4, and Low yields 34.1, so the High - Low difference in estimates is -0.7.
* If dep\_start = 1, High ed\_start yields 35.7, and Low yields 38.6, so the High - Low difference in estimates is -2.9.
* Estimated effect of ed\_start **depends** on dep\_start.

## Effect of dep\_start = 1 vs. 0?

| Est. bmi\_end | ed\_start = High | ed\_start = Low |
| --- | --- | --- |
| dep\_start = 1 | 35.7 | 38.6 |
| dep\_start = 0 | 33.4 | 34.1 |

* If ed\_start is High, then dep\_start = 1 yields 35.7, and 0 yields 33.4, so the difference in estimated bmi\_end is 2.3.
* If ed\_start is Low, then dep\_start = 1 yields 38.6, and 0 yields 34.1, so the difference in estimated bmi\_end is 4.5.
* Estimated effect of dep\_start **depends** on ed\_start.

## ANOVA table for withint model

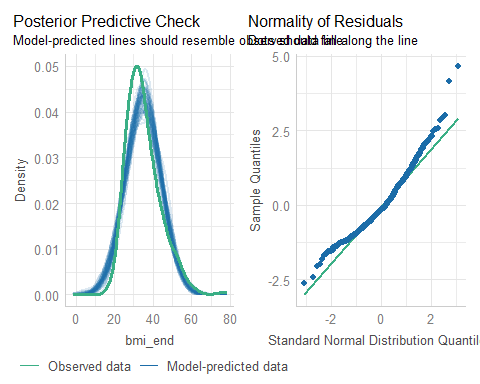
anova(withint)

Analysis of Variance Table  
  
Response: bmi\_end  
 Df Sum Sq Mean Sq F value Pr(>F)   
dep\_start 1 1303 1302.59 17.1299 4.152e-05 \*\*\*  
ed\_start 1 280 280.48 3.6884 0.05541 .   
dep\_start:ed\_start 1 133 133.25 1.7523 0.18625   
Residuals 460 34979 76.04   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* Total SS = 1303 + 280 + 133 + 34979 = 36696
* % explained = (1303 + 280 + 133) / 36696 = 0.047, or 4.7%
* Is this a big improvement over the mainfx model?

## Checking model withint

set.seed(202410291)  
check\_model(withint, check = c("pp\_check", "qq"), detrend = FALSE)



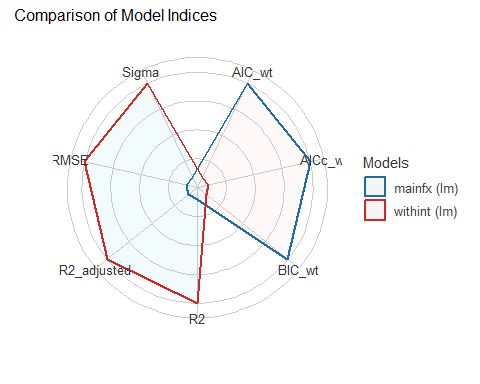
## Compare peformance of our two models?

compare\_performance(mainfx, withint, rank = TRUE)

# Comparison of Model Performance Indices  
  
Name | Model | R2 | R2 (adj.) | RMSE | Sigma | AIC weights | AICc weights | BIC weights | Performance-Score  
------------------------------------------------------------------------------------------------------------------  
withint | lm | 0.047 | 0.041 | 8.683 | 8.720 | 0.471 | 0.465 | 0.101 | 57.14%  
mainfx | lm | 0.043 | 0.039 | 8.699 | 8.727 | 0.529 | 0.535 | 0.899 | 42.86%

## Plot to compare performance indicators

plot(compare\_performance(mainfx, withint))



## Do we have a substantial interaction?

* Interaction plot shows slightly non-parallel lines.
* % of variation explained by the interaction term is about 0.4% of the variation in bmi\_end, overall.
* p value for the interaction term is pretty large ( = 0.19)
* Is there a substantial improvement in the fit with the interaction?
  + In-sample performance indices are split.

What do you think?

# Two-Factor Analysis of Variance: Example 2

## Example 2

Suppose that we now want to understand the impact on BMI at the end of the period, of:

* insur\_start (Medicare, Commercial or Medicaid), and
* ed\_start (Low or High)

and it’s possible that the impact of insurance on BMI may depend on the home neighborhood’s educational attainment, and vice versa (i.e. insurance and education may **interact**.)

## Build an interaction plot

Calculate and store group means.

ex2\_means <- dm18 |>  
 group\_by(insur\_start, ed\_start) |>  
 summarize(mean\_bmi = mean(bmi\_end))

`summarise()` has grouped output by 'insur\_start'. You can override using the  
`.groups` argument.

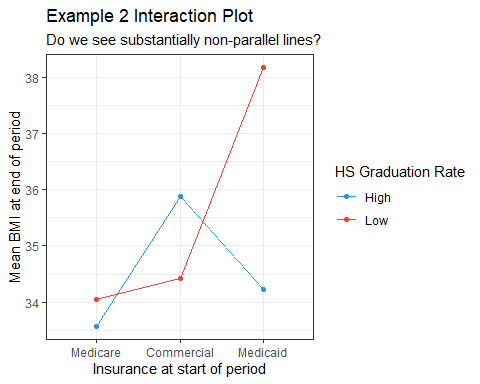
ex2\_means

# A tibble: 6 × 3  
# Groups: insur\_start [3]  
 insur\_start ed\_start mean\_bmi  
 <fct> <fct> <dbl>  
1 Medicare High 33.6  
2 Medicare Low 34.1  
3 Commercial High 35.9  
4 Commercial Low 34.4  
5 Medicaid High 34.2  
6 Medicaid Low 38.2

## How strong is the interaction?

Now, we’ll plot these means, and look for a substantial interaction (non-parallel lines.)

ggplot(ex2\_means, aes(x = insur\_start, y = mean\_bmi)) +  
 geom\_line(aes(group = ed\_start, color = ed\_start)) +  
 geom\_point(aes(color = ed\_start)) +  
 scale\_color\_material() +  
 labs(title = "Example 2 Interaction Plot",  
 subtitle = "Do we see substantially non-parallel lines?",  
 y = "Mean BMI at end of period",   
 x = "Insurance at start of period",   
 color = "HS Graduation Rate")



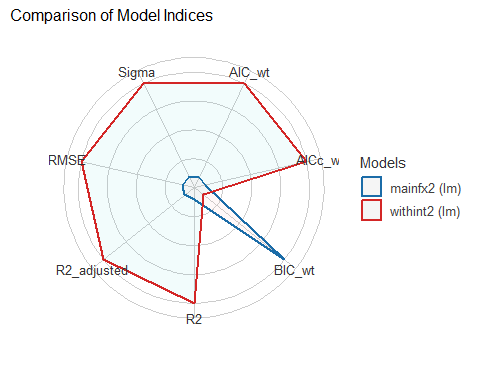
## Example 2: Two-Way ANOVA models

mainfx2 <- lm(bmi\_end ~ insur\_start + ed\_start, data = dm18)  
withint2 <- lm(bmi\_end ~ insur\_start \* ed\_start, data = dm18)  
  
compare\_performance(mainfx2, withint2, rank = TRUE)

# Comparison of Model Performance Indices  
  
Name | Model | R2 | R2 (adj.) | RMSE | Sigma | AIC weights | AICc weights | BIC weights | Performance-Score  
-------------------------------------------------------------------------------------------------------------------  
withint2 | lm | 0.037 | 0.027 | 8.727 | 8.784 | 0.758 | 0.747 | 0.048 | 85.71%  
mainfx2 | lm | 0.024 | 0.018 | 8.786 | 8.824 | 0.242 | 0.253 | 0.952 | 14.29%

## Plotting Performance Indices

plot(compare\_performance(mainfx2, withint2))



## Interpret mainfx2 estimates?

model\_parameters(mainfx2, ci = 0.90)

Parameter | Coefficient | SE | 90% CI | t(460) | p  
--------------------------------------------------------------------------------  
(Intercept) | 32.85 | 0.78 | [31.56, 34.13] | 42.02 | < .001  
insur start [Commercial] | 1.44 | 1.30 | [-0.70, 3.57] | 1.11 | 0.269   
insur start [Medicaid] | 2.40 | 0.88 | [ 0.94, 3.85] | 2.72 | 0.007   
ed start [Low] | 1.79 | 0.82 | [ 0.44, 3.15] | 2.18 | 0.030

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

| Est. bmi\_end | Medicare | Commercial | Medicaid |
| --- | --- | --- | --- |
| ed\_start High | 32.9 | 34.3 | 35.3 |
| ed\_start Low | 34.7 | 36.1 | 37.1 |

## Specify withint2 equation?

model\_parameters(withint2, ci = 0.90)

Parameter | Coefficient | SE | 90% CI | t(458) | p  
-------------------------------------------------------------------------------------------------  
(Intercept) | 33.57 | 0.95 | [32.01, 35.13] | 35.44 | < .001  
insur start [Commercial] | 2.31 | 1.84 | [-0.72, 5.34] | 1.26 | 0.210   
insur start [Medicaid] | 0.66 | 1.26 | [-1.42, 2.74] | 0.53 | 0.599   
ed start [Low] | 0.48 | 1.27 | [-1.61, 2.58] | 0.38 | 0.704   
insur start [Commercial] × ed start [Low] | -1.94 | 2.58 | [-6.20, 2.32] | -0.75 | 0.452   
insur start [Medicaid] × ed start [Low] | 3.44 | 1.76 | [ 0.55, 6.34] | 1.96 | 0.051

Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed  
 using a Wald t-distribution approximation.

## withint2 model equation

$$
bmi\_{end} = 33.6 + 2.3 Commercial + 0.7 Medicaid + 0.5 (ed\_{start} = Low) \\ - 1.9 Commercial \times (ed\_{start} = Low) + 3.4 Medicaid \times (ed\_{start} = Low)
$$

| insur\_start | ed\_start | Estimated bmi\_end |
| --- | --- | --- |
| Medicare | High | 33.6 |
| Commercial | High | 33.6 + 2.3 = 35.9 |
| Medicaid | High | 33.6 + 0.7 = 34.3 |
| Medicare | Low | 33.6 + 0.5 = 34.1 |
| Commercial | Low | 33.6 + 2.3 + 0.5 - 1.9 = 34.5 |
| Medicaid | Low | 33.6 + 0.7 + 0.5 + 3.4 = 38.2 |

## ANOVA table for main effects model

anova(mainfx2)

Analysis of Variance Table  
  
Response: bmi\_end  
 Df Sum Sq Mean Sq F value Pr(>F)   
insur\_start 2 510 254.82 3.2727 0.03879 \*  
ed\_start 1 369 369.45 4.7450 0.02989 \*  
Residuals 460 35817 77.86   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### % of Total SS explained by the two factors?

* Total SS = 510 + 369 + 35817 = 36696
* So % explained = (510 + 369) / 36696 = 0.024, or 2.4%

## ANOVA table for interaction model

anova(withint2)

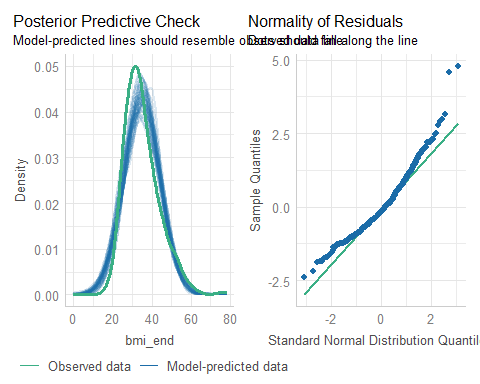
Analysis of Variance Table  
  
Response: bmi\_end  
 Df Sum Sq Mean Sq F value Pr(>F)   
insur\_start 2 510 254.82 3.3029 0.03765 \*  
ed\_start 1 369 369.45 4.7888 0.02915 \*  
insur\_start:ed\_start 2 482 240.88 3.1223 0.04500 \*  
Residuals 458 35335 77.15   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### % of Total SS explained by the model?

* Total SS = 510 + 369 + 482 + 35335 = 36696
* So % explained = (510 + 369 + 482) / 36696 = 0.037, or 3.7%

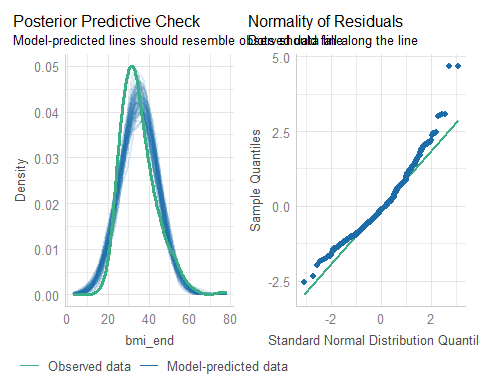
## Checking model mainfx2

set.seed(202410292)  
check\_model(mainfx2, check = c("pp\_check", "qq"), detrend = FALSE)



## Checking model withint2

set.seed(202410293)  
check\_model(withint2, check = c("pp\_check", "qq"), detrend = FALSE)



## Do we have a substantial interaction?

Example 2

* Interaction plot shows substantially non-parallel lines.
* Interaction term explains about 1.3% of the variation in bmi\_end, overall.
* p value for the interaction term is fairly small ( = 0.045)
* Substantial improvement in the fit with interaction?
  + 6/7 in-sample performance indices prefer interaction

What do you think?

# Add a covariate?

## What if we added a covariate?

dm18a <- dm464 |>  
 mutate(ed\_start = factor(ifelse(nedlev\_start > 78, "High", "Low")),  
 dep\_start = factor(depdiag\_start)) |>  
 mutate(bmi\_c = center(bmi\_start)) |>  
 select(id\_code, bmi\_end, dep\_start, ed\_start, insur\_start,   
 bmi\_c, bmi\_start)  
  
dm18a |> reframe(lovedist(bmi\_start))

# A tibble: 1 × 10  
 n miss mean sd med mad min q25 q75 max  
 <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 464 0 35.2 8.82 33.4 8.01 17.3 28.9 40.4 75.1

dm18a |> reframe(lovedist(bmi\_c))

# A tibble: 1 × 10  
 n miss mean sd med mad min q25 q75 max  
 <int> <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 464 0 2.30e-15 8.82 -1.84 8.01 -17.9 -6.34 5.21 39.9

## Analysis of Covariance Model

fit3 <- lm(bmi\_end ~ bmi\_c + dep\_start \* insur\_start, data = dm18a)  
  
anova(fit3)

Analysis of Variance Table  
  
Response: bmi\_end  
 Df Sum Sq Mean Sq F value Pr(>F)   
bmi\_c 1 34213 34213 6413.8007 < 2e-16 \*\*\*  
dep\_start 1 8 8 1.4107 0.23556   
insur\_start 2 29 14 2.7165 0.06718 .   
dep\_start:insur\_start 2 9 4 0.8135 0.44394   
Residuals 457 2438 5   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Proportion of Variation Explained

eta\_squared(fit3)

# Effect Size for ANOVA (Type I)  
  
Parameter | Eta2 (partial) | 95% CI  
-----------------------------------------------------  
bmi\_c | 0.93 | [0.93, 1.00]  
dep\_start | 3.08e-03 | [0.00, 1.00]  
insur\_start | 0.01 | [0.00, 1.00]  
dep\_start:insur\_start | 3.55e-03 | [0.00, 1.00]  
  
- One-sided CIs: upper bound fixed at [1.00].

## Drop Interaction?

fit4 <- lm(bmi\_end ~ bmi\_c + dep\_start + insur\_start, data = dm18a)  
  
anova(fit4)

Analysis of Variance Table  
  
Response: bmi\_end  
 Df Sum Sq Mean Sq F value Pr(>F)   
bmi\_c 1 34213 34213 6419.0169 < 2e-16 \*\*\*  
dep\_start 1 8 8 1.4119 0.23536   
insur\_start 2 29 14 2.7187 0.06702 .   
Residuals 459 2446 5   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Drop Depression?

fit5 <- lm(bmi\_end ~ bmi\_c + insur\_start, data = dm18a)  
  
anova(fit5)

Analysis of Variance Table  
  
Response: bmi\_end  
 Df Sum Sq Mean Sq F value Pr(>F)   
bmi\_c 1 34213 34213 6414.4440 < 2e-16 \*\*\*  
insur\_start 2 29 15 2.7587 0.06442 .   
Residuals 460 2453 5   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

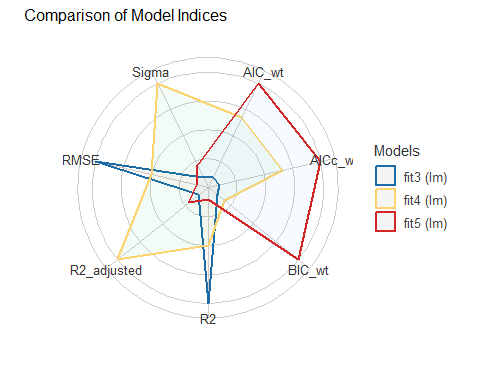
## Compare fit3, fit4, fit5

compare\_performance(fit3, fit4, fit5, rank = TRUE)

# Comparison of Model Performance Indices  
  
Name | Model | R2 | R2 (adj.) | RMSE | Sigma | AIC weights | AICc weights | BIC weights | Performance-Score  
---------------------------------------------------------------------------------------------------------------  
fit4 | lm | 0.933 | 0.933 | 2.296 | 2.309 | 0.370 | 0.368 | 0.083 | 60.74%  
fit5 | lm | 0.933 | 0.933 | 2.300 | 2.309 | 0.515 | 0.526 | 0.916 | 46.38%  
fit3 | lm | 0.934 | 0.933 | 2.292 | 2.310 | 0.114 | 0.106 | 4.09e-04 | 28.57%

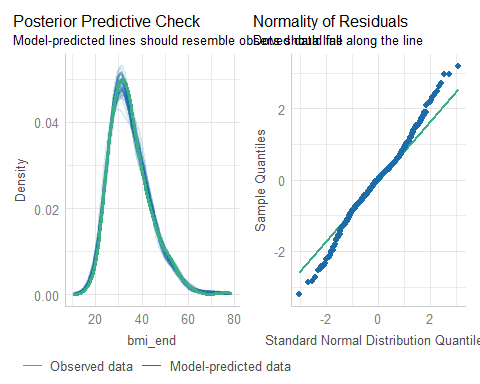
## Compare fit3, fit4, fit5

plot(compare\_performance(fit3, fit4, fit5))



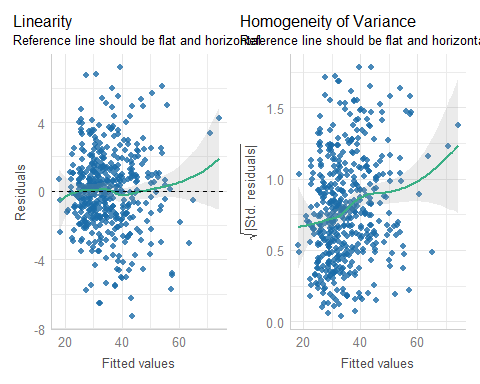
## Check Model fit4 (1/3)

check\_model(fit4, check = c("pp\_check", "qq"), detrend = FALSE)



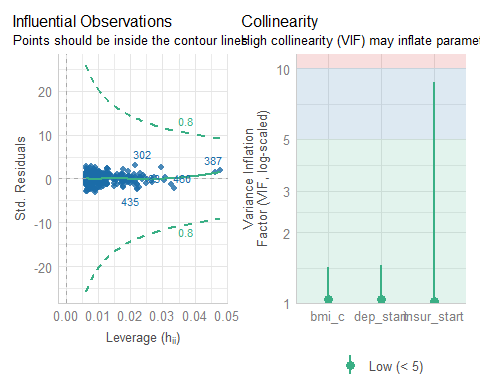
## Check Model fit4 (2/3)

check\_model(fit4, check = c("linearity", "homogeneity"))



## Check Model fit4 (3/3)

check\_model(fit4, check = c("outliers", "vif"))



## Session Information

xfun::session\_info()

R version 4.4.1 (2024-06-14 ucrt)  
Platform: x86\_64-w64-mingw32/x64  
Running under: Windows 11 x64 (build 22631)  
  
Locale:  
 LC\_COLLATE=English\_United States.utf8   
 LC\_CTYPE=English\_United States.utf8   
 LC\_MONETARY=English\_United States.utf8  
 LC\_NUMERIC=C   
 LC\_TIME=English\_United States.utf8   
  
Package version:  
 askpass\_1.2.1 backports\_1.5.0 base64enc\_0.1.3   
 bayestestR\_0.15.0 bit\_4.5.0 bit64\_4.5.2   
 blob\_1.2.4 broom\_1.0.7 bslib\_0.8.0   
 cachem\_1.1.0 callr\_3.7.6 cellranger\_1.1.0   
 cli\_3.6.3 clipr\_0.8.0 coda\_0.19-4.1   
 codetools\_0.2-20 colorspace\_2.1-1 compiler\_4.4.1   
 conflicted\_1.2.0 correlation\_0.8.6 cpp11\_0.5.0   
 crayon\_1.5.3 curl\_5.2.3 data.table\_1.16.2   
 datasets\_4.4.1 datawizard\_0.13.0 DBI\_1.2.3   
 dbplyr\_2.5.0 digest\_0.6.37 dplyr\_1.1.4   
 dtplyr\_1.3.1 easystats\_0.7.3 effectsize\_0.8.9   
 emmeans\_1.10.5 estimability\_1.5.1 evaluate\_1.0.1   
 fansi\_1.0.6 farver\_2.1.2 fastmap\_1.2.0   
 fontawesome\_0.5.2 forcats\_1.0.0 fs\_1.6.4   
 gargle\_1.5.2 generics\_0.1.3 ggplot2\_3.5.1   
 ggrepel\_0.9.6 glue\_1.8.0 googledrive\_2.1.1   
 googlesheets4\_1.1.1 graphics\_4.4.1 grDevices\_4.4.1   
 grid\_4.4.1 gtable\_0.3.6 haven\_2.5.4   
 highr\_0.11 hms\_1.1.3 htmltools\_0.5.8.1   
 httr\_1.4.7 ids\_1.0.1 insight\_0.20.5   
 isoband\_0.2.7 janitor\_2.2.0 jquerylib\_0.1.4   
 jsonlite\_1.8.9 knitr\_1.48 labeling\_0.4.3   
 lattice\_0.22-6 lifecycle\_1.0.4 lubridate\_1.9.3   
 magrittr\_2.0.3 MASS\_7.3-61 Matrix\_1.7-0   
 memoise\_2.0.1 methods\_4.4.1 mgcv\_1.9-1   
 mime\_0.12 modelbased\_0.8.9 modelr\_0.1.11   
 multcomp\_1.4-26 munsell\_0.5.1 mvtnorm\_1.3-1   
 nlme\_3.1-164 numDeriv\_2016.8.1.1 openssl\_2.2.2   
 parameters\_0.23.0 patchwork\_1.3.0 performance\_0.12.4   
 pillar\_1.9.0 pkgconfig\_2.0.3 prettyunits\_1.2.0   
 processx\_3.8.4 progress\_1.2.3 ps\_1.8.0   
 purrr\_1.0.2 R6\_2.5.1 ragg\_1.3.3   
 rappdirs\_0.3.3 RColorBrewer\_1.1.3 Rcpp\_1.0.13   
 readr\_2.1.5 readxl\_1.4.3 rematch\_2.0.0   
 rematch2\_2.1.2 report\_0.5.9 reprex\_2.1.1   
 rlang\_1.1.4 rmarkdown\_2.28 rstudioapi\_0.17.1   
 rvest\_1.0.4 sandwich\_3.1-1 sass\_0.4.9   
 scales\_1.3.0 see\_0.9.0 selectr\_0.4.2   
 snakecase\_0.11.1 splines\_4.4.1 stats\_4.4.1   
 stringi\_1.8.4 stringr\_1.5.1 survival\_3.7-0   
 sys\_3.4.3 systemfonts\_1.1.0 textshaping\_0.4.0   
 TH.data\_1.1-2 tibble\_3.2.1 tidyr\_1.3.1   
 tidyselect\_1.2.1 tidyverse\_2.0.0 timechange\_0.3.0   
 tinytex\_0.53 tools\_4.4.1 tzdb\_0.4.0   
 utf8\_1.2.4 utils\_4.4.1 uuid\_1.2.1   
 vctrs\_0.6.5 viridisLite\_0.4.2 vroom\_1.6.5   
 withr\_3.0.2 xfun\_0.48 xml2\_1.3.6   
 xtable\_1.8-4 yaml\_2.3.10 zoo\_1.8-12

1. The assessment of linearity and non-constant variance used by check\_model() are best used when we have at least one quantitative predictor in our model. [↑](#footnote-ref-38)