Homework #5

Dustin Leatherman

February 23, 2019

1

Is this dataset balanced?

This dataset is balanced. Each Age and Process variable has 10 observations

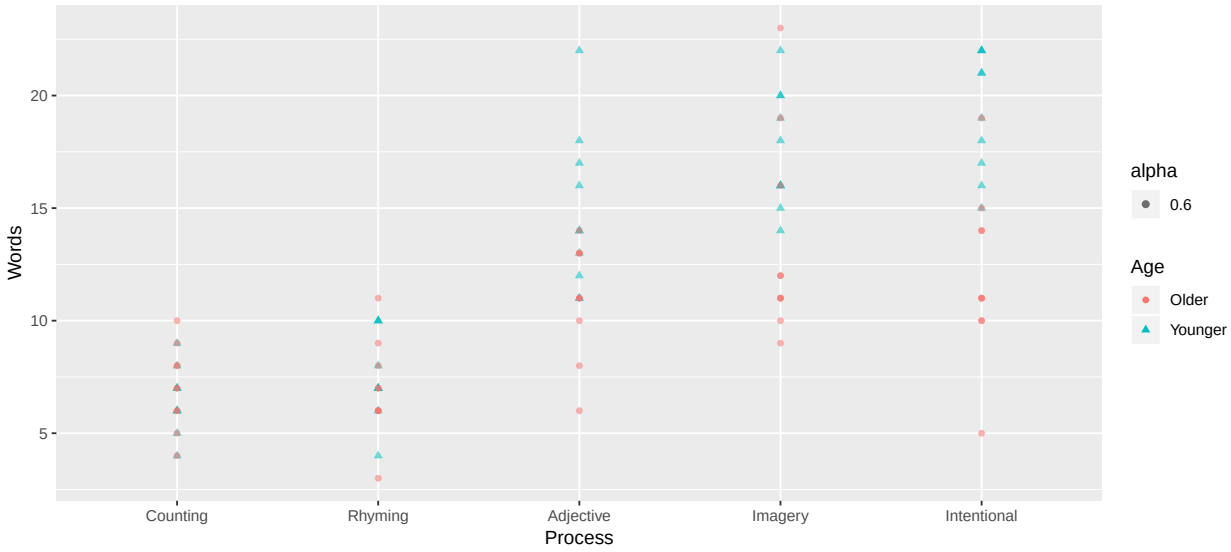
```
words %>%
  group_by(Process, Age) %>%
  summarise_at("Words", funs(n=n())) %>%
  kable %>%
    kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

Process	Age	n
Counting	Older	10
Counting	Younger	10
Rhyming	Older	10
Rhyming	Younger	10
Adjective	Older	10
Adjective	Younger	10
Imagery	Older	10
Imagery	Younger	10
Intentional	Older	10
Intentional	Younger	10

2

Create a plot of the raw data

```
words %>%
  ggplot(aes(x = Process, y = Words, color = Age, shape = Age, alpha = 0.6)) +
    geom_point()
```



3

Produce a two-way table of sample averages, along with row and column averages. Using the table you produce: - Under the saturated model, what is the estimated difference in mean number of words recalled between the younger and older groups in the adjective treatment?

```
words %>%
    group_by(Process, Age) %>%
    # get mean for each group
    summarize_at("Words", funs(Mean=mean)) %>%
    # Break into two-tab formula
    unstack(form = Mean ~ Process) %>%
    # calculate the Total column the total across rows
    mutate(Total = rowSums(.)) %>%
    # Calculate the total row. This isn't pretty but it works :/ Would love to know a better way
    rbind(list(
      Adjective = sum(.[1]),
      Counting = sum(.[2]),
      Imagery = sum(.[3]),
      Intentional = sum(.[4]),
      Rhyming = sum(.[5]),
      Total = sum(.[6]))
    ) %>%
    # Create a column and convert it to the row name for output
    mutate(i1 = c("Older","Younger", "Total")) %>%
    column_to_rownames("i1") %>%
  kable(
    caption = "Average Word Count for number of Words Recalled"
  ) %>%
      kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position") %>%
      column_spec(1, bold = T, color = "black", border_right = T) %>%
      row_spec(3, color = "black", bold = T) %>%
      column_spec(7, color = "black", bold = T)
```

Table 1: Average Word Count for number of Words Recalled

	Counting	Rhyming	Adjective	Imagery	Intentional	Total
Older	7.0	6.9	11.0	13.4	12.0	50.3
Younger	6.5	7.6	14.8	17.6	19.3	65.8
Total	**14.5**	**31.3**	**13.5**	**25.8**	**31.0**	**116.1**

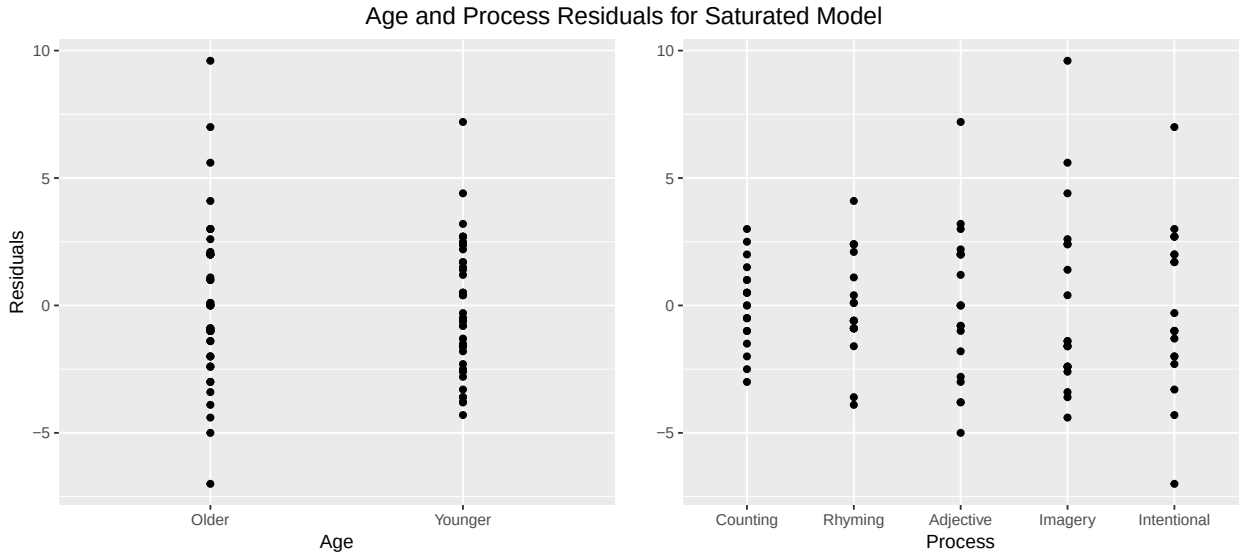
65.8 - 50.3 = 15.5 = ~16

It is estimated that the younger group would recall an average of 16 words more than the older group after controlling for the word recalled.

4

Fit the saturated model and examine the residuals. Is there any evidence of a need for transformation?

```
model.full <- lm(Words ~ Age * Process, words)
age.res <- qplot(Age, .resid, data = model.full) + ylab("Residuals")
process.res <- qplot(Process, .resid, data = model.full) + ylab("")

grid.arrange(age.res, process.res,
             ncol = 2,
             widths = c(1, 1),
             top = textGrob("Age and Process Residuals for Saturated Model",
                            gp=gpar(fontsize=14,font=1),just=c("center")))
```



There are no distinct patterns in the residual charts so there is no evidence that transformations are required.

5

Run an extra sum of squares F-test to compare the saturated model to the additive model. Is there evidence against the simpler additive model?

```r
model.add <- lm(Words ~ Age + Process, words)

tidy(anova(model.add, model.full)) %>%
  kable(
    caption = "Extra Sum of Squares F-Test for Saturated vs Additive Model"
  ) %>%
    kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

Table 2: Extra Sum of Squares F-Test for Saturated vs Additive Model

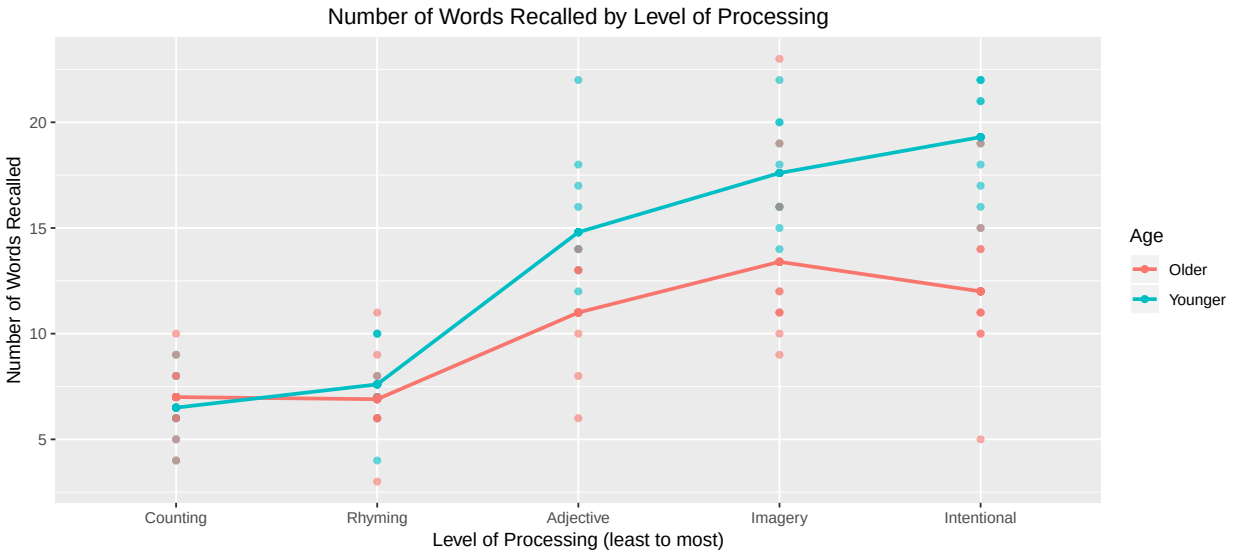
res.df	rss	df	sumsq	statistic	p.value
94	912.6	NA	NA	NA	NA
90	722.3	4	190.3	5.927938	0.0002793

There is convincing evidence that the full model is a better model to use than the additive model (Extra Sum of Squares F-Test. p-value = 0.0003)

6

Using the saturated model, produce a plot of the estimated mean responses.

```r
model.full %>%
  # sort the x-axis by the number of words
  ggplot(aes(x = reorder(Process, Words), y = Words, group = Age, color = Age)) +
  # plot original data
  geom_point(alpha = 0.6) +
  # Add a best-fit line
  geom_line(aes(Process, .fitted), size = 1) +
  # Highlight the fitted points so we can see them clearly
  geom_point(aes(Process, .fitted)) +
    ylab("Number of Words Recalled") +
    xlab("Level of Processing (least to most)") +
    ggtitle("Number of Words Recalled by Level of Processing") +
      theme(plot.title = element_text(hjust = 0.5))
```

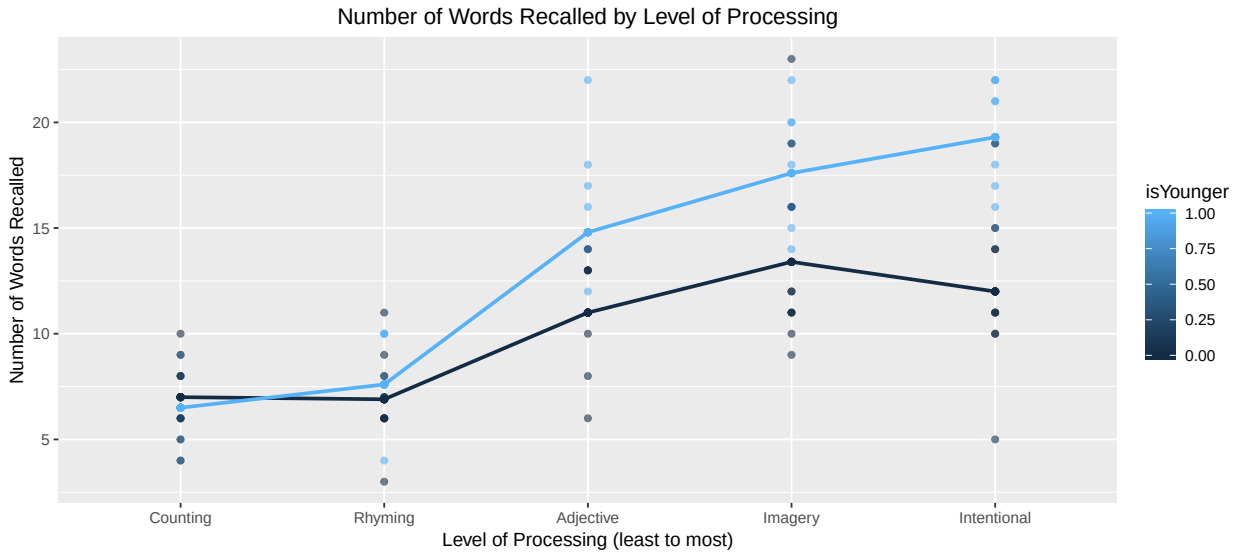


7

```
# Make 'Adjective' the baseline
words$Process <- relevel(words$Process, ref = "Adjective")

# Create indicator variables for each Process Level
words$isYounger <- ifelse(words$Age == "Younger", 1, 0)
words$isRhyming <- ifelse(words$Process == "Rhyming", 1, 0)
words$isIntentional <- ifelse(words$Process == "Intentional", 1, 0)
words$isImagery <- ifelse(words$Process == "Imagery", 1, 0)
words$isCounting <- ifelse(words$Process == "Counting", 1, 0)
words$isAdjective <- ifelse(words$Process == "Adjective", 1, 0)

model.full.reparam <- lm(Words ~ Process + isAdjective:isYounger + isCounting:isYounger + isImagery:isY

tidy(model.full.reparam) %>%
  kable %>%
    kable_styling(full_width = T, bootstrap_options = "striped", latex_options = "hold_position")
```

term	estimate	std.error	statistic	p.value
(Intercept)	11.0	0.8958547	12.2787776	0.0000000
ProcessCounting	-4.0	1.2669298	-3.1572389	0.0021676
ProcessRhyming	-4.1	1.2669298	-3.2361698	0.0016959
ProcessImagery	2.4	1.2669298	1.8943433	0.0613907
ProcessIntentional	1.0	1.2669298	0.7893097	0.4320055
isAdjective:isYounger	3.8	1.2669298	2.9993769	0.0034984
isYounger:isCounting	-0.5	1.2669298	-0.3946549	0.6940313
isYounger:isImagery	4.2	1.2669298	3.3151008	0.0013216
isYounger:isIntentional	7.3	1.2669298	5.7619609	0.0000001
isYounger:isRhyming	0.7	1.2669298	0.5525168	0.5819640

```r
model.full.reparam %>%
  # sort the x-axis by the number of words
  ggplot(aes(x = reorder(Process, Words), y = Words, group = isYounger, color = isYounger)) +
    # plot original data
    geom_point(alpha = 0.6) +
    # Add a best-fit line
    geom_line(aes(Process, .fitted), size = 1) +
    # Highlight the fitted points so we can see them clearly
    geom_point(aes(Process, .fitted)) +
      ylab("Number of Words Recalled") +
      xlab("Level of Processing (least to most)") +
      ggtitle("Number of Words Recalled by Level of Processing") +
      theme(plot.title = element_text(hjust = 0.5))
```



Using 'Intentional' as an example

$\beta_0 + \beta_3 + \beta_8$ - $(\beta_0 + \beta_3)$

$\rightarrow \beta_8$

β_8 represents the difference in mean words between older and younger people for the Process Level 'Intentional'. Since all of the process levels are represented similarly, this explanation can be applied to each.

There is convincing evidence that young people have an advantage on Intentional, Imagery, and Adjective Processing (p-values =1.15e-07, 0.00132, and 0.0035 respectively).