Typing Speed

First, define the mean structure.

```
n     <- 50  # choose the n
x0   <-  0   # population mean for x
y0   <-  0   # population mean for x
v_x0 <-  1   # variance around x
v_y0 <-  1   # variance around y
cov  <- -.8  # covariance for the variances

# the next three lines of code simply combine the terms, above
mu     <- c(x0, y0)
sigma  <- matrix(c(v_x0, cov,
                   cov, v_y0), ncol = 2)
set.seed(1)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tibble)
library(purrr)
library(ggplot2)
library(tidyr)
library(stringr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v lubridate 1.9.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readr)
library(forcats)

m <-
  MASS::mvrnorm(n, mu, sigma) %>%
  data.frame() %>%
```

```
  set_names("x0", "y0") %>%
  arrange(x0) %>%
  mutate(i = 1:n) %>%
  expand(nesting(i, x0, y0),
         j = 1:n)
```

Second, define the residual structure.

```
# note how these values initially place the epsilons in a standardized metric
sigma  <- matrix(c(v_x0, .3,
                   .3, v_y0), ncol = 2)
set.seed(1)
r <-
  MASS::mvrnorm(n * n, mu, sigma) %>%
  data.frame() %>%
  set_names("e_x", "e_y") %>%
  # you do not need this step.
  # it's something I experimented with to rescale
  # the residual variances to a workable level for the plots.
  mutate_all(.funs = ~. * .25)
```

Combine the two data structures and (optionally) save.

```
typingSpeed <-
  bind_cols(m, r) %>%
  mutate(typing_speed = x0 + e_x,
         num_typos = y0 + e_y)
```

If you have the CSV file on disk, load it instead of regenerating the data.

```
# library(readr)
# typingSpeed <- read_csv("typingSpeed.csv")
# typingSpeed$participant <- as.factor(typingSpeed$participant)
```

Now the fun part.

Suppose we are interested in the relationship between typing speed (i.e., number of words typed per minute) and the percentage of typos that are made. If we look at the cross-sectional relationship (i.e., the population level), we may well find a negative relationship, in that people who type faster make fewer mistakes (this may be reflective of the fact that people with better developed typing skills and more experience both type faster and make fewer mistakes).

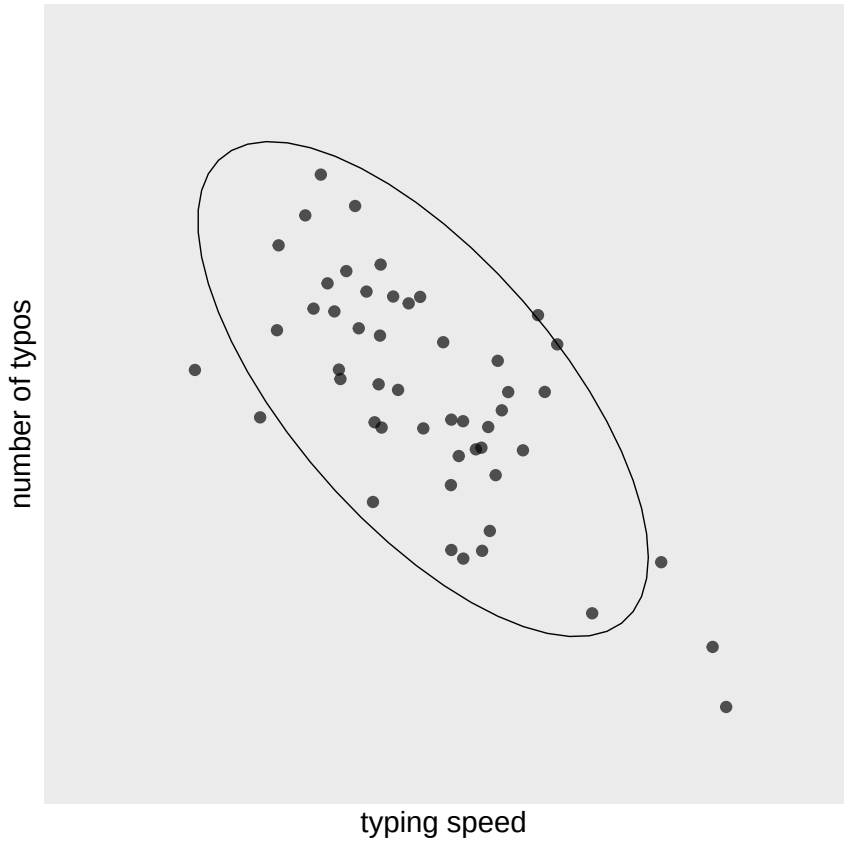
```
library(tidyverse)

typingSpeed %>%
  filter(j == 1) %>%

  ggplot(aes(x = typing_speed, y = num_typos)) +
  geom_point(alpha = 2/3) +
  stat_ellipse(size = 1/4) +
  scale_x_continuous("typing speed", breaks = NULL, limits = c(-3, 3)) +
  scale_y_continuous("number of typos", breaks = NULL, limits = c(-3, 3)) +
  coord_equal() +
  theme(panel.grid = element_blank())
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
```

```
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Model the relationship

```
m1 = lm(typing_speed ~ num_typos, data = typingSpeed)
summary(m1)
```

```
##
## Call:
## lm(formula = typing_speed ~ num_typos, data = typingSpeed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39598 -0.42371  0.03279  0.43660  2.02526
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.09850    0.01324  -7.441 1.37e-13 ***
## num_typos   -0.63823    0.01484 -42.999  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6605 on 2498 degrees of freedom
## Multiple R-squared:  0.4253, Adjusted R-squared:  0.4251
## F-statistic:  1849 on 1 and 2498 DF,  p-value: < 2.2e-16
```

The simulated data I just showed were from 50 participants at one measurement occasion. However, the full data set contains 50 measurement occasions for each of the 50 participants. In the next plot, we'll show all 50 measurement occasions for just 5 participants. The n is reduced in the plot to avoid cluttering.

The points are colored by participant. The gray points in each data cloud are the participant-level means. Although we still see a clear negative relationship between participants, we now also see a mild positive relationship within participants.

```
typingSpeed %>%
  filter(i %in% c(1, 10, 25, 47, 50)) %>%
  mutate(i = factor(i)) %>%

  ggplot() +
  geom_point(aes(x = typing_speed, y = num_typos, color = i),
             size = 1/3) +
  stat_ellipse(aes(x = typing_speed, y = num_typos, color = i),
               size = 1/5) +
  geom_point(data = typingSpeed %>%
               filter(j == 1 &
                        i %in% c(1, 10, 25, 47, 50)),
             aes(x = x0, y = y0),
             size = 2, color = "grey50") +
  scale_color_viridis_d(option = "B", begin = .25, end = .85) +
  scale_x_continuous("typing speed", breaks = NULL, limits = c(-3, 3)) +
  scale_y_continuous("number of typos", breaks = NULL, limits = c(-3, 3)) +
  coord_equal() +
  theme(panel.grid = element_blank(),
        legend.position = "none")
```