

# Handling experimental data with R

## Day 5

SALOS 2025

Alena Witzlack-Makarevich

## Getting ready

---

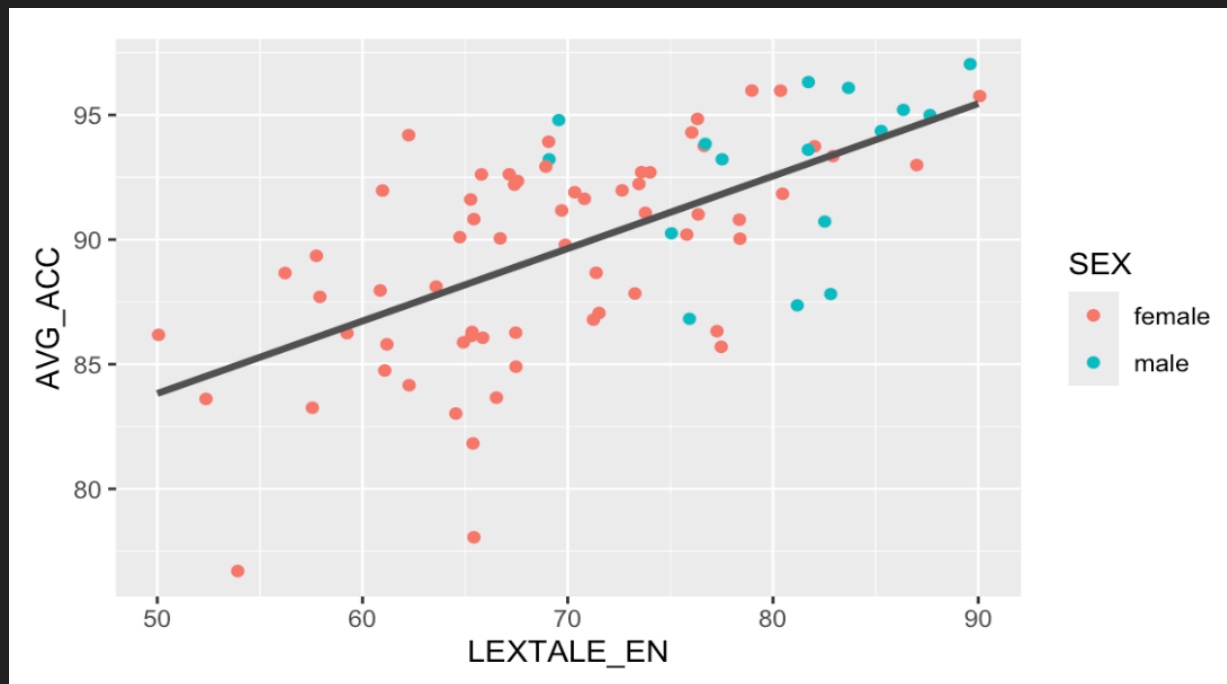
- You have the GECO R notebook open
- It is super clean
- You loaded tidyverse with `library("tidyverse")`
- You have the `GECO` dataset loaded into R
- You derived and cleaned the data frame `participants`

**Simple linear regression:**  
**one numeric predictor variable,**  
**one numeric response variable**

## Simple linear regression (cont.)

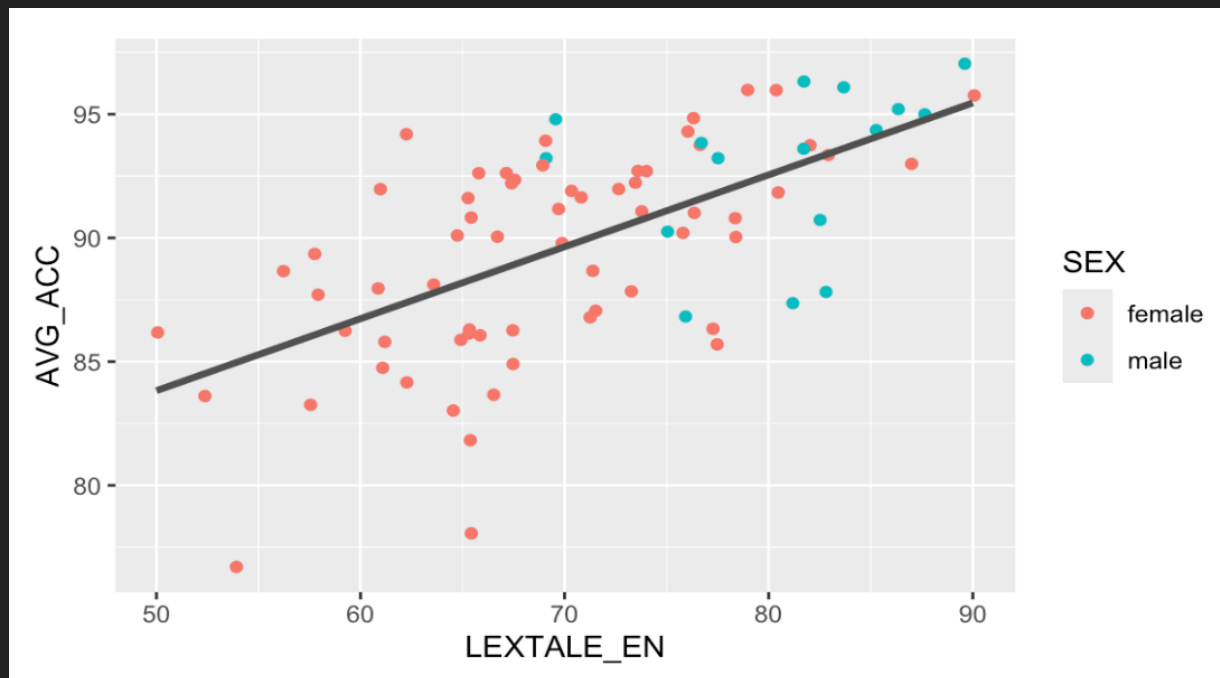
## The line of best fit

- The simplest kind of model we can **fit** is a straight line (linear model)
- Called **the line of best fit** or **a linear regression line**



## The line of best fit

- Fitted with  
`lm(AVG_ACC ~ LEXTALE_EN, data = participants)`



## Linear model with `lm()`

Call:

```
lm(formula = AVG_ACC ~ LEXTALE_EN, data = .)
```

Residuals:

|  | Min      | 1Q      | Median | 3Q     | Max    |
|--|----------|---------|--------|--------|--------|
|  | -10.1872 | -2.1872 | 0.4512 | 2.5416 | 6.5397 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )    |
|-------------|----------|------------|---------|-------------|
| (Intercept) | 69.28720 | 3.00195    | 23.081  | < 2e-16 *** |
| LEXTALE_EN  | 0.29077  |            |         |             |

This gives us:  $y = 0.29 * x + 69.3$

How do you interpret this number?

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.366 on 78 degrees of freedom

Multiple R-squared: 0.3828, Adjusted R-squared: 0.3748

F-statistic: 48.37 on 1 and 78 DF, p-value: 9.607e-10

## Linear model with `lm()`

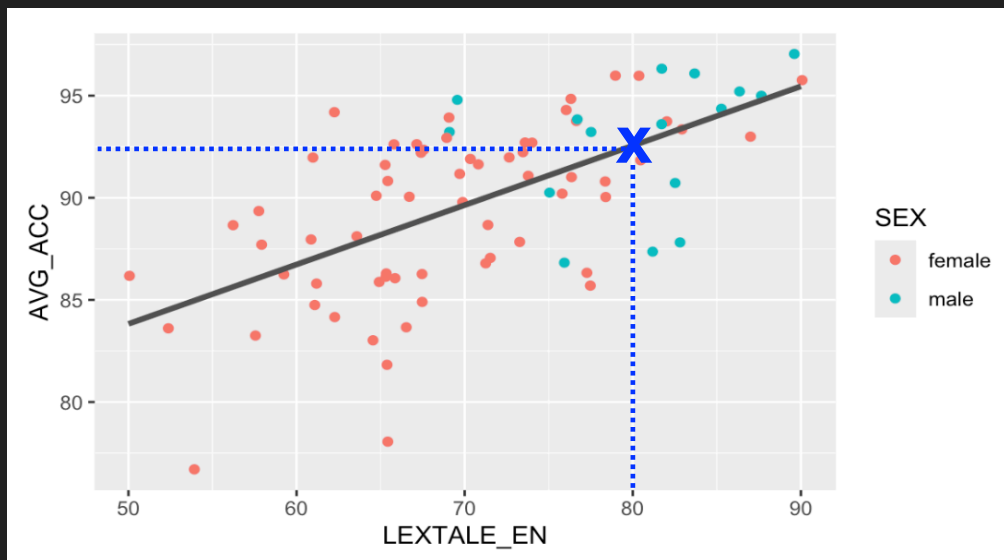
- Described with this equation:

$$y = 0.29 * x + 69.3$$

- It can be used to predict future observations
- E.g. if `LEXTALE_EN = 80`, the model predicts `AVG_ACC = 92.5`

Coefficients:

|             | Estimate |
|-------------|----------|
| (Intercept) | 69.28720 |
| LEXTALE_EN  | 0.29077  |



## Linear model with `lm()`

Call:

```
lm(formula = AVG_ACC ~ LEXTALE_EN, data = .)
```

Residuals:

| Min      | 1Q      | Median | 3Q     | Max    |
|----------|---------|--------|--------|--------|
| -10.1872 | -2.1872 | 0.4512 | 2.5416 | 6.5397 |

Coefficients:

|             | Estimate | Std. Error | t-value | Pr(> t )     |
|-------------|----------|------------|---------|--------------|
| (Intercept) | 69.28720 | 3.00195    | 23.081  | < 2e-16      |
| LEXTALE_EN  | 0.29077  | 0.04181    | 6.955   | 9.61e-10 *** |

If  $p < 0.05$ ,  $H_0$  of no effect of LEXTALE\_EN can be rejected

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.366 on 78 degrees of freedom

Multiple R-squared: 0.3828, Adjusted R-squared: 0.3748

F-statistic: 48.37 on 1 and 78 DF, p-value: 9.607e-10

## Linear model with `lm()`

Call:

```
lm(formula = AVG_ACC ~ LEXTALE_EN, data = .)
```

Residuals:

| Min      | 1Q      | Median | 3Q     | Max    |
|----------|---------|--------|--------|--------|
| -10.1872 | -2.1872 | 0.4512 | 2.5416 | 6.5397 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t )    |
|-------------|----------|------------|---------|-------------|
| (Intercept) | 69.28720 | 3.00195    | 23.081  | < 2e-16 *** |

LEX *p*-value is the probability of observing a given test statistics  
--- (here *F*-ratio) if  $H_0$  were true.

Sig Since  $p < 0.05$  is very low (very unlikely), we conclude  
Res that the model is significant.

Multiple R-squared: 0.3828, Adjusted R-squared: 0.3748

F-statistic: 48.37 on 1 and 78 DF, *p*-value: 9.607e-10

## Linear model with `lm()`

Call:

```
lm(formula = AVG_ACC ~ LEXTALE_EN, data = .)
```

Residuals:

| Min      | 1Q      | Median | 3Q     | Max    |
|----------|---------|--------|--------|--------|
| -10.1872 | -2.1872 | 0.4512 | 2.5416 | 6.5397 |

Coefficients:

|             | Estimate | Std. |
|-------------|----------|------|
| (Intercept) | 69.28720 | 3.   |
| LEXTALE_EN  | 0.29077  | 0.   |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The model explains approximately 38% of the variability in AVG\_ACC.

That is, 38% of the differences in AVG\_ACC across observations can be accounted for by LEXTALE\_EN.

Explained variance

366 on 78 degrees of freedom

Multiple R-squared: 0.3828, Adjusted R-squared: 0.3748

F-statistic: 48.37 on 1 and 78 DF, p-value: 9.607e-10

## Adding a regression line

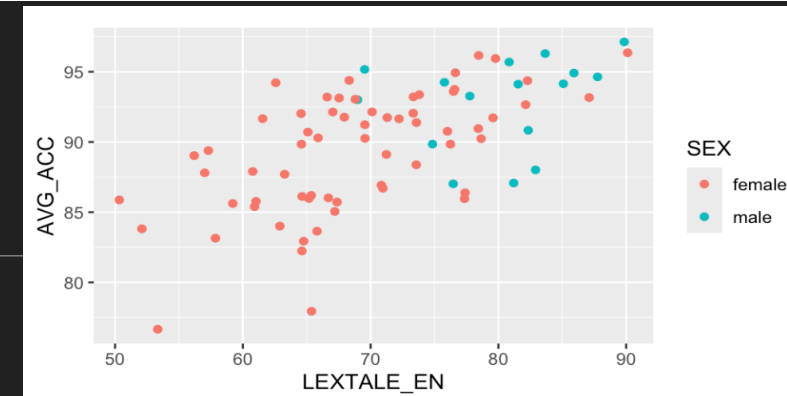
- We combine two geoms:  
`geom_jitter` (or `geom_point`) and  
`geom_smooth()`

- Like this:

```
participants %>%
```

```
  ggplot(aes(x = LEXTALE_EN, y = AVG_ACC, color = SEX)) +  
  geom_jitter() +  
  geom_smooth()
```

- What does the first attempt give you? Why?
- What do you want to improve?

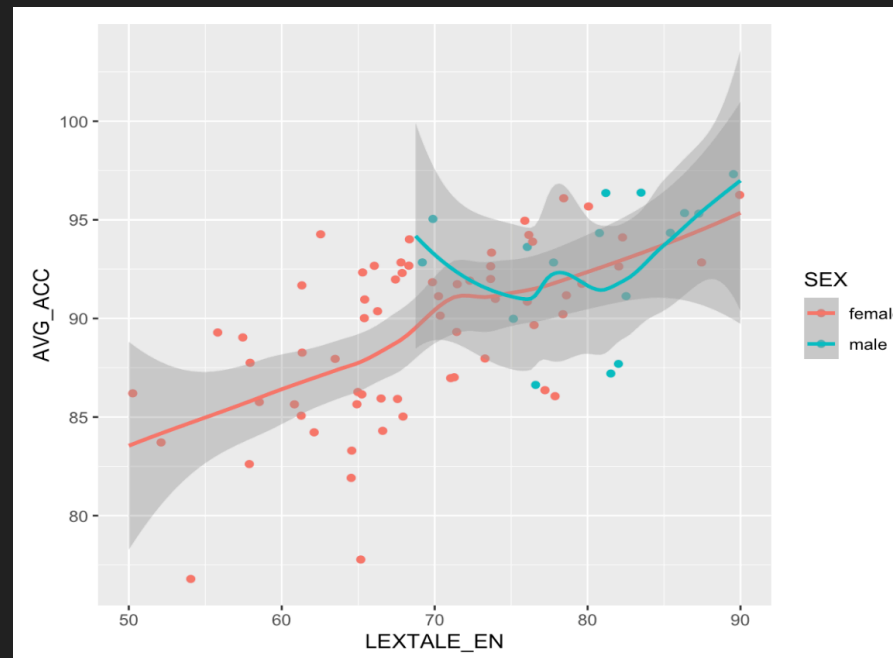


`geom_line()` is already  
“occupied” for line charts

## Adding a regression line

- Two regression lines: one for **male** and one for **female**. Why?  
participants %>%

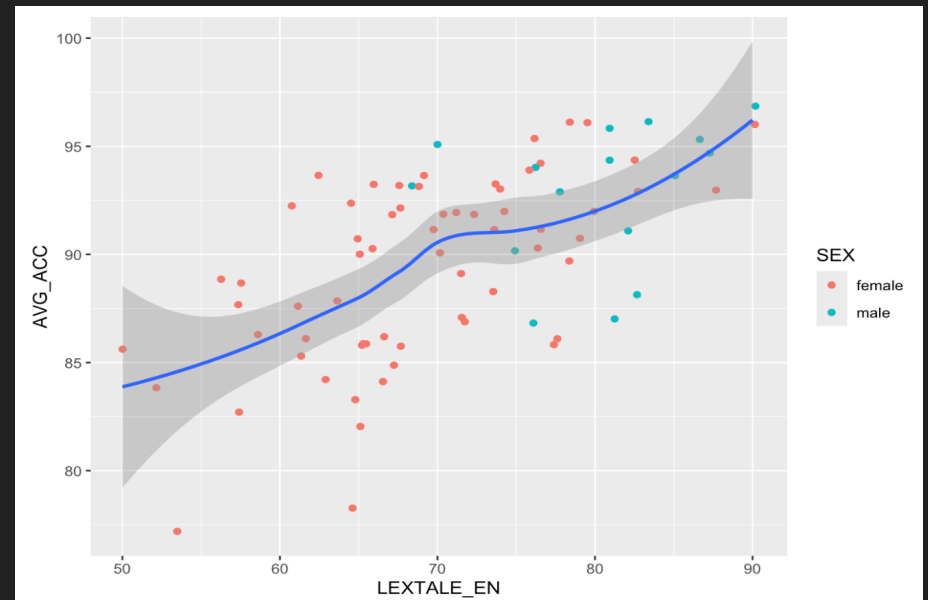
```
ggplot(aes(x = LEXTALE_EN, y = AVG_ACC, color = SEX)) +  
  geom_jitter() +  
  geom_smooth()
```



## Adding a regression line

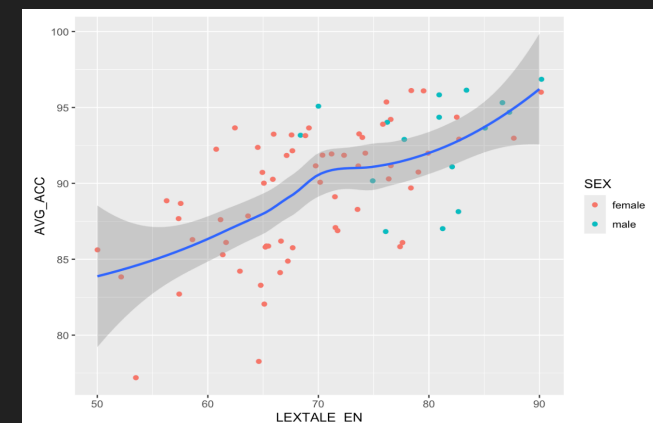
- We want `color = SEX` to apply only to `geom_point()`:  
`participants %>%`

```
ggplot(aes(x = LEXTALE_EN, y = AVG_ACC)) +  
  geom_jitter(aes(color = SEX)) +  
  geom_smooth()
```



## Adding a regression line

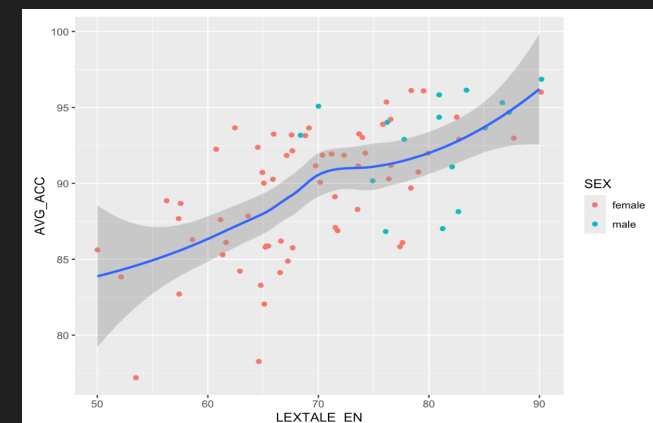
- We have a smoothing line (**smoother**): shows in more detail what the trends look like
- aka **local regression** or moving regression
- This can be especially helpful when trying to understand regressions
- Most common methods are LOESS (locally estimated scatterplot smoothing) and LOWESS (locally weighted scatterplot smoothing)
- Both pronounced /'loʊɛs/



## Adding a regression line

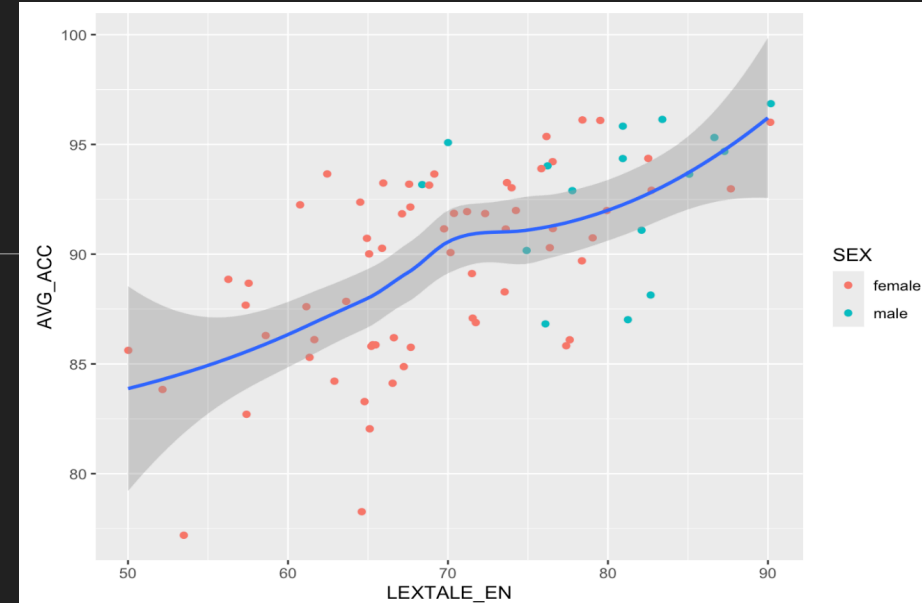
- We have a smoothing line (**smoother**): shows in more detail what the trends look like
- aka **local regression** or moving regression
- This can be especially helpful when trying to understand regressions
- Most common methods are LOESS (locally estimated scatterplot smoothing) and LOWESS (locally weighted scatterplot smoothing)
- Both pronounced /'loʊɛs/
- Great, but I do want a line!

```
geom_smooth(method = "lm")
```

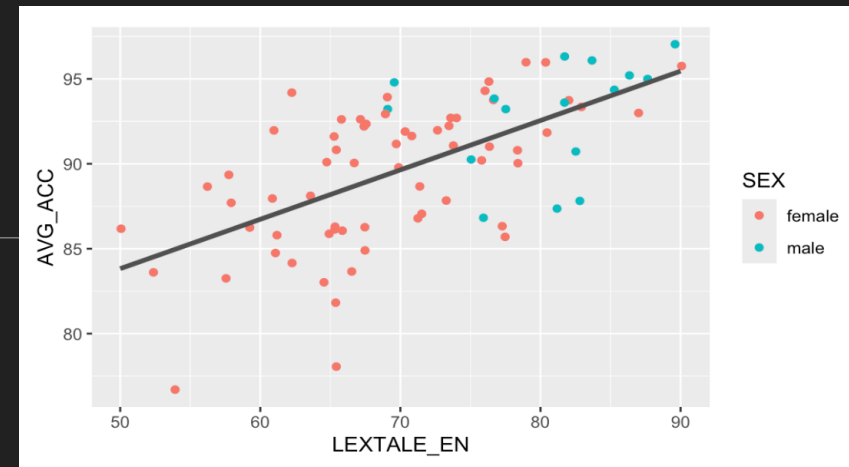


## Standard error

- What the grey ribbon?
- The **standard error (se)** shows how much the actual data points tend to vary above or below the predicted line
- It gives us an idea of the uncertainty in our predictions
- If the standard error is small, the data points are close to the line
- Get rid of it with  
`geom_smooth(method = "lm", se = FALSE)`



## And you can do more



- `geom_smooth(method = "lm", se = FALSE,  
color = "grey33", size = 1.2)`

## Facets are the best!

---

- Try **facets**:

```
geom_smooth(method = "lm", se = FALSE) +  
facet_grid(. ~ SEX)
```

Theme

Coordinates

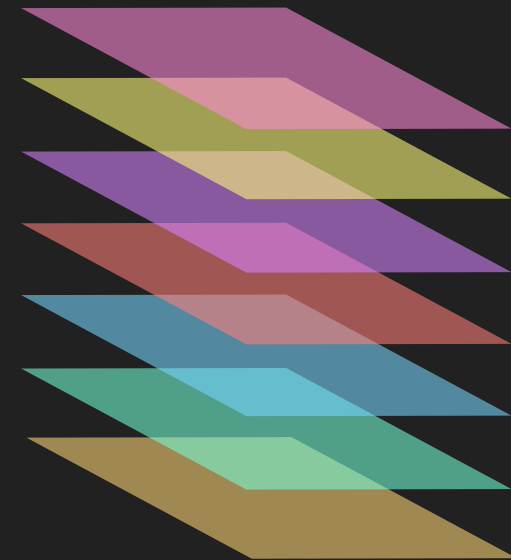
Statistics

**Facets**

Geometries

Aesthetics

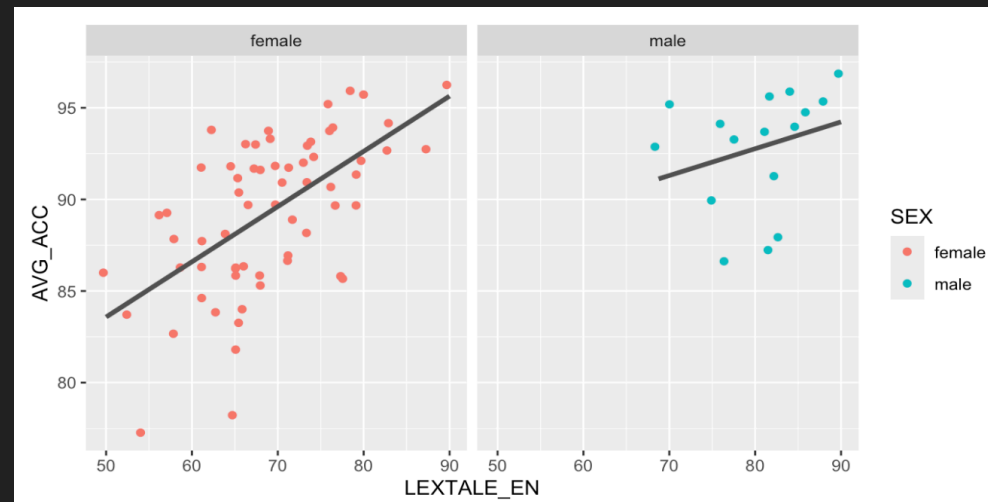
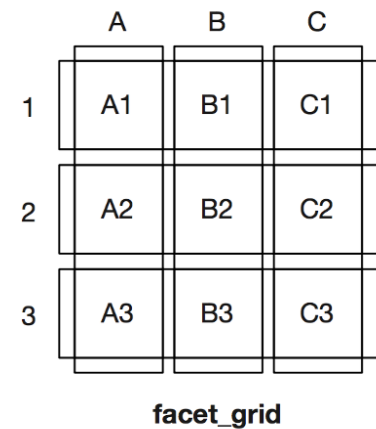
**Data**



## Facets are the best!

- Facet grid tells you how to add subplots on the x and y-axes: `facet_grid(y ~ x)`
- Here we add subplots along the x axis by SEX and no subplot on the y-axis

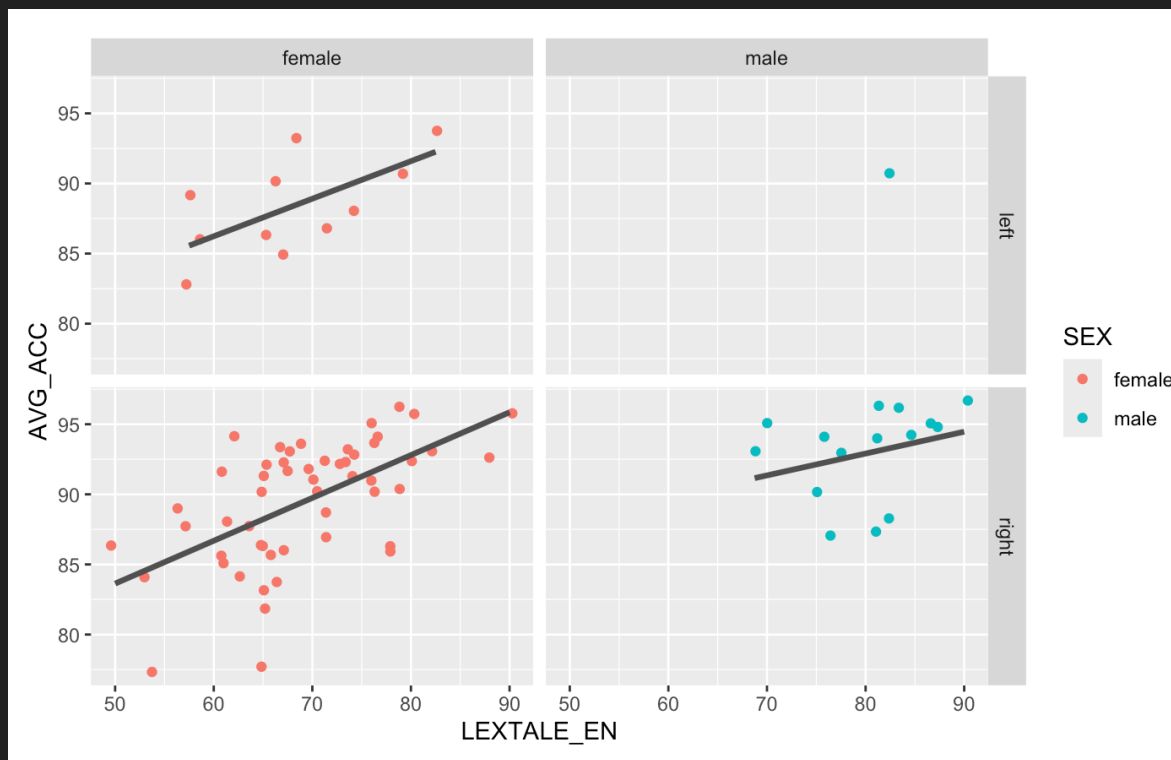
`facet_grid(. ~ SEX)`



## Facets are the best!

- You can combine two facet variables:

```
facet_grid(HANDEDNESS ~ SEX)
```



**Simple linear regression:  
other predictors**

## Looking for other correlations

- So **LEXTALE\_EN** is a good predictor of **AVG\_ACC**
- But is it the best predictor?
- Can another predictor explain more variance?
- Can a combination of predictors explain more variance?

A tibble: 81 × 7

| <b>PPNR</b><br><dbl> | <b>HANDEDNESS</b><br><chr> | <b>SEX</b><br><chr> | <b>AGE</b><br><dbl> | <b>AVG_ACC</b><br><dbl> | <b>LEXTALE_EN</b><br><chr> | <b>LEXTALE_DU</b><br><chr> |
|----------------------|----------------------------|---------------------|---------------------|-------------------------|----------------------------|----------------------------|
| 1                    | right                      | female              | 20                  | 92                      | 80                         | 85                         |
| 17                   | right                      | female              | 18                  | 92                      | 71.25                      | 86.25                      |
| 25                   | right                      | female              | 18                  | 86                      | 50                         | 88.75                      |
| 33                   | right                      | female              | 18                  | 96                      | 78.75                      | 87.5                       |
| 41                   | riah                       | male                | 18                  | 93                      | 68.75                      | 83.75                      |

## LEXTALE\_DU as a predictor

---

- a. Produce a scatterplot of `AVG_ACC` vs `LEXTALE_DU`
- b. Add a regression line to the plot with:  
`geom_smooth(method = "lm")`
- c. Estimate the correlation coefficient
- d. And then calculate it with `cor()`
- e. Estimate the slope and then model the relationship
- f. Compare the outcomes of the two models:  
does the model with `LEXTALE_EN` explains more variance  
than the model with `LEXTALE_DU`?

## LEXTALE\_DU as a predictor

---

a. Produce a scatterplot of AVG\_ACC vs LEXTALE\_DU

```
participants %>%
```

```
  ggplot(aes(x = LEXTALE_DU, y = AVG_ACC)) +  
  geom_jitter(aes(color = SEX)) +  
  geom_smooth(method = "lm", se = FALSE
```

## LEXTALE\_DU as a predictor

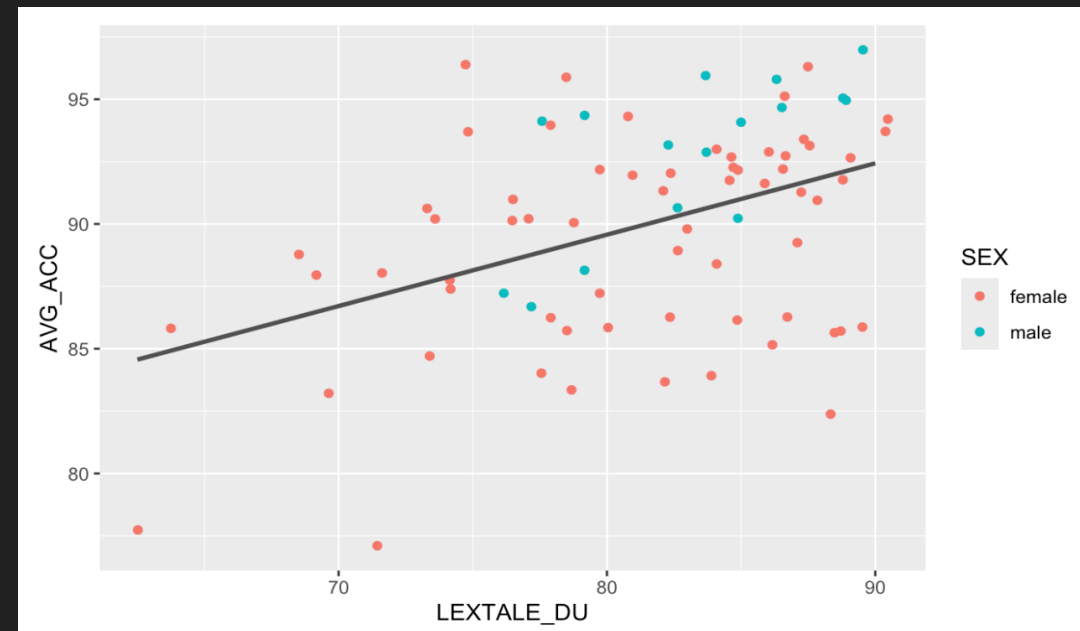
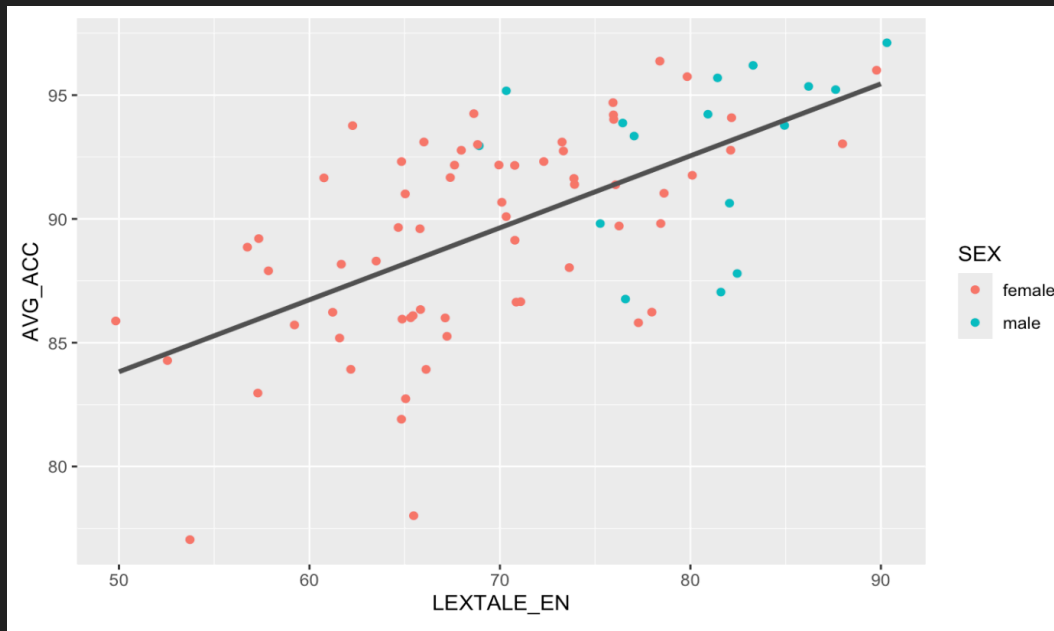
a. Which correlation coefficient is higher?

LEXTALE\_EN

vs

LEXTALE\_DU

b. What actually matter? Tell your neighbor



## LEXTALE\_DU as a predictor

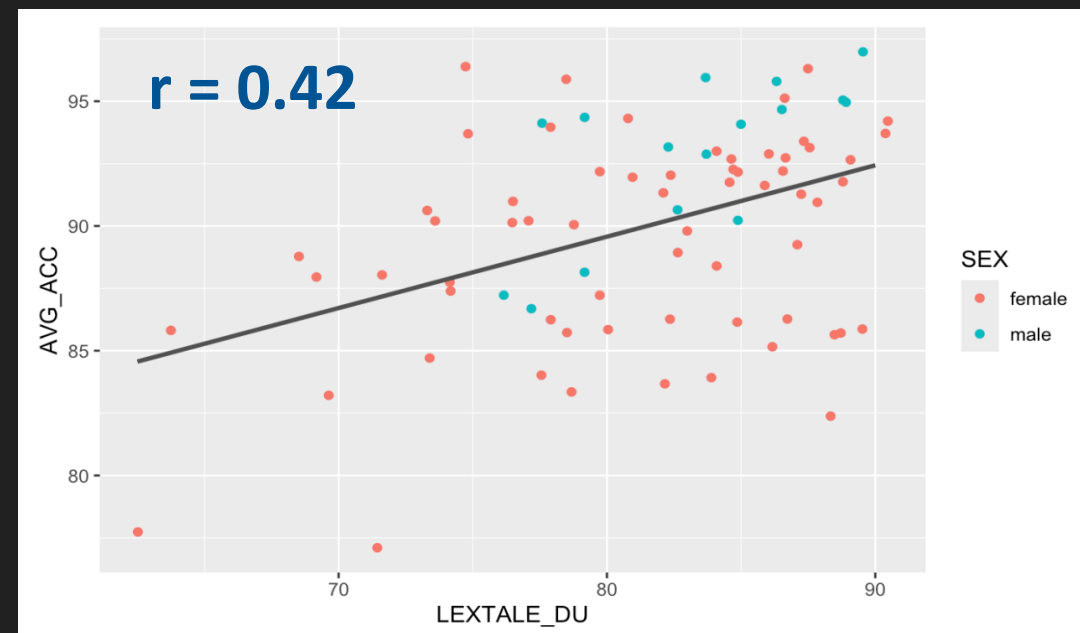
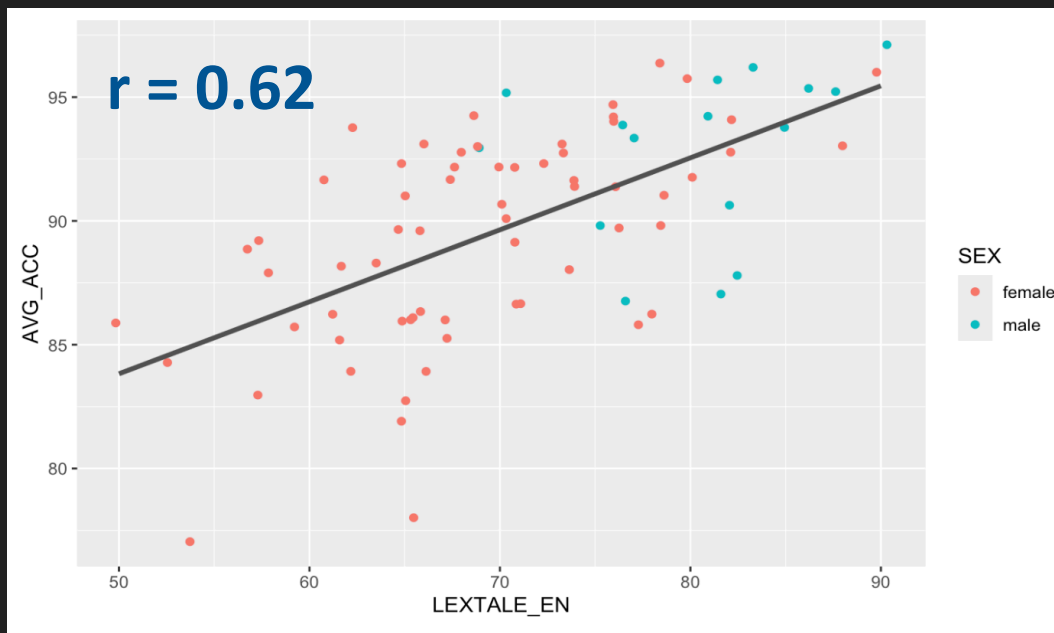
a. Which correlation coefficient is higher?

LEXTALE\_EN

vs

LEXTALE\_DU

b. What actually matter? Tell your neighbor



## LEXTALE\_EN vs LEXTALE\_DU

Compare the outcomes of the two models: does the model with **LEXTALE\_EN** explain more variance than the model with **LEXTALE\_DU**?

```
Call:
lm(formula = AVG_ACC ~ LEXTALE_EN, data = participants)

Residuals:
    Min       1Q   Median       3Q      Max
-10.1872  -2.1872   0.4512   2.5416   6.5397

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.28720    3.00195   23.081 < 2e-16 ***
LEXTALE_EN   0.29077    0.04181    6.955 9.61e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.366 on 78 degrees of freedom
Multiple R-squared:  0.3828,    Adjusted R-squared:  0.3748
F-statistic: 48.37 on 1 and 78 DF,  p-value: 9.607e-10
```

```
Call:
lm(formula = AVG_ACC ~ LEXTALE_DU, data = participants)

Residuals:
    Min       1Q   Median       3Q      Max
-10.0743  -2.6620   0.9656   2.4474   7.8597

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.68199    5.64590   11.811 < 2e-16 ***
LEXTALE_DU   0.28611    0.06907    4.142 8.64e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.879 on 78 degrees of freedom
Multiple R-squared:  0.1803,    Adjusted R-squared:  0.1698
F-statistic: 17.16 on 1 and 78 DF,  p-value: 8.644e-05
```

**Simple linear regression:**  
**nominal/categorical predictors**

## SEX as a categorical predictors

---

- a. Simple linear regression also allows binary nominal/categorical predictors, such as ...
- b. Try to model it with `lm()`
- c. Compare the outcomes of this model with the previous two models (`LEXTALE_EN` and `LEXTALE_DU` as predictors)
- d. A standard way to **visualize** this relation is a boxplot or a bar plot with means and error bars.  
Produce a boxplot to compare the distributions.  
Add `geom_jitter()` atop, adjust alpha (e.g. `alpha = 0.3`)

## Categorical predictors

---

a. Simple linear regression also allows binary nominal/categorical predictors, such as ...

b. Try to model it with `lm()`

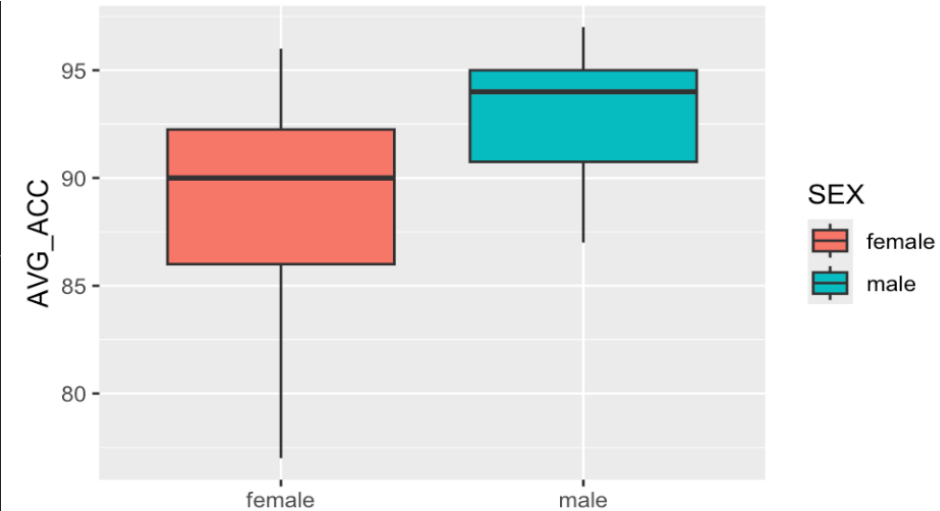
```
model_SEX <- lm(AVG_ACC ~ SEX, data =participants)
summary(model_SEX)
```

c. Compare the outcomes of this model with the previous two models (`LEXTALE_EN` and `LEXTALE_DU` as predictors)

d. A standard way to visualize this relation is a **boxplot** or a **bar plot with means and error bars** (criticized for being visually limited or even misleading)

## SEX as a categorical predictors

- How is the presentation of the results different from quantitative predictors?
- How do you interpret the slope?
- How much variation does the model explain?



Call:

```
lm(formula = AVG_ACC ~ SEX, data = participants)
```

Residuals:

| Min      | 1Q      | Median | 3Q     | Max    |
|----------|---------|--------|--------|--------|
| -12.2969 | -3.2969 | 0.7031 | 2.7031 | 6.7031 |

Coefficients:

|             | Estimate | Std. Error | t value | Pr(> t ) |     |
|-------------|----------|------------|---------|----------|-----|
| (Intercept) | 89.2969  | 0.5051     | 176.776 | < 2e-16  | *** |
| SEXmale     | 3.5156   | 1.1295     | 3.112   | 0.00259  | **  |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.'

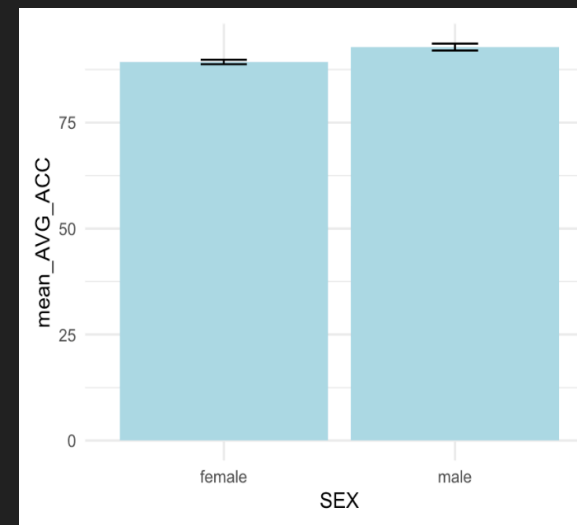
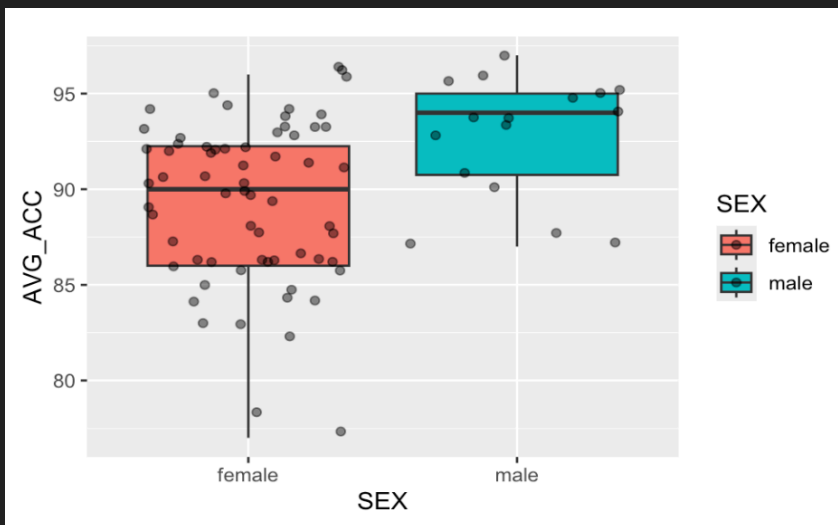
Residual standard error: 4.041 on 78 degrees of freedom

Multiple R-squared: 0.1105, Adjusted R-squared:

F-statistic: 9.687 on 1 and 78 DF, p-value: 0.002593

## SEX as a categorical predictors

- A standard way to visualize this relation is a **boxplot** or a **bar plot with means and error bars** (criticized for being visually limited or even misleading)



SE here = uncertainty in sample mean per group

# **Multiple regression**

## Multiple predictors

---

- It seems that on their own **LEXTALE\_EN**, **LEXTALE\_DU**, and **SEX** are all good predictors of **AVG\_ACC**
- Can we combine their predictive power and explain more of the variation in **AVG\_ACC**?
- Or maybe whatever seems to be explained by **SEX** is actually already explained by e.g. **LEXTALE\_EN**?
- What if we add **AGE** and **HANDEDNESS**?
- Multiple regression lets you examine how several predictors together influence an outcome!

## Multiple regression

---

- The equation:

$$y = b_1*x + b_2*x + b_3*x + a$$

- The function:

```
m_all <- lm(AVG_ACC ~ AGE + SEX + ... + LEXTALE_DU,  
data = participants)  
summary(m_all)
```

- Discuss the output with your neighbor and interpret the results.
- What matters for **AVG\_ACC**?

## Multiple regression

Call:

```
lm(formula = AVG_ACC ~ AGE + SEX + HANDEDNESS + LEXTALE_EN +  
    LEXTALE_DU, data = participants)
```

Residuals:

| Min     | 1Q      | Median | 3Q     | Max    |
|---------|---------|--------|--------|--------|
| -7.0821 | -2.1829 | 0.4851 | 2.3099 | 5.8149 |

Coefficients:

|                 | Estimate | Std. Error | t value | Pr(> t ) |     |
|-----------------|----------|------------|---------|----------|-----|
| (Intercept)     | 60.15879 | 7.34655    | 8.189   | 5.69e-12 | *** |
| AGE             | -0.36912 | 0.32593    | -1.133  | 0.261074 |     |
| SEXmale         | 0.18664  | 1.02718    | 0.182   | 0.856311 |     |
| HANDEDNESSright | 0.75677  | 0.99598    | 0.760   | 0.449773 |     |
| LEXTALE_EN      | 0.25835  | 0.04595    | 5.622   | 3.16e-07 | *** |
| LEXTALE_DU      | 0.21550  | 0.05804    | 3.713   | 0.000394 | *** |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.149 on 74 degrees of freedom

Multiple R-squared: 0.4876, Adjusted R-squared: 0.453

F-statistic: 14.09 on 5 and 74 DF, p-value: 1.134e-09

**Next challenge**

## What's next?

---

- Get back to the original GECO dataset
- Explore some possible predictors of RT
- Model the relationships
- Try multivariate regression
- But you'll actually need to use a different type of model:
- RT is **not normally distributed**, you will need to log-transform or use a different model
- There are **random effects** (PPNR and related categories):  
this will allow each participant to have their own baseline (intercept) and account for non-independence of repeated measures