stats_practice

Nicole Jimenez

2022-10-26

```
##set up libraries
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(tidyverse)
```

```
## ── Attaching packages
## ─────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.5
## ✓ tibble  3.1.8     ✓ stringr 1.4.1
## ✓ tidyr   1.2.1     ✓ forcats 0.5.2
## ✓ readr   2.1.3
## ── Conflicts ──────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(readr)
library(ggplot2)
library(data.table)
```

```
##
## Attaching package: 'data.table'
##
## The following object is masked from 'package:purrr':
##
##     transpose
##
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

Load Data

```
##Load feature data - species
Cer_vaginal_species_feature <-
  read_csv("C:/Users/nicolejimenez/OneDrive - University of Arizona/Desktop/Atopobium/cervical_c
ancer_NAU/cervical_cancer_NAU_level_7_taxa_table.csv")
```

```
## Rows: 297 Columns: 99
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (1): #NAME
## dbl (98): Sample_1, Sample_2, Sample_3, Sample_4, Sample_5, Sample_6, Sample...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(Cer_vaginal_species_feature)
```

```
#Load metadata check to make sure it is csv not txt!
Cer_AV_vag_meta <- read_csv("C:/Users/nicolejimenez/OneDrive - University of Arizona/Desktop/Ato
pobium/cervical_cancer_NAU/Cervical_cancer_NAU_metadata_level_7.csv")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 1032290 Columns: 21
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (6): #NAME, Atopobium_vaginae_types, T_disease_state, T_inflammation, T...
## dbl (15): Atopobium vaginae, Atopobium deltae, Atopobium vaginae_A, Atopobia...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(Cer_AV_vag_meta)
```

Statistical Question: Does the number of Atopobiaceae species types affect the genital inflammatory score?

Null hypothesis: The number of Atopobiaceae species types does not affect genital inflammatory score Alternate hypothesis: the number of Atopobiacea species types does affect the genital inflammatory score

```
model <- lm(T_infl_score_flt ~ Atopobiaceae_types, data = Cer_AV_vag_meta)
summary(model)$coef
```

```
##                     Estimate Std. Error  t value    Pr(>|t|)
## (Intercept)        1.3686552  0.3029662 4.517517 0.000017815
## Atopobiaceae_types 0.2847985  0.1589580 1.791660 0.076338322
```

genital inflammation = 1.369 + 0.284*Atopobiaceae_species

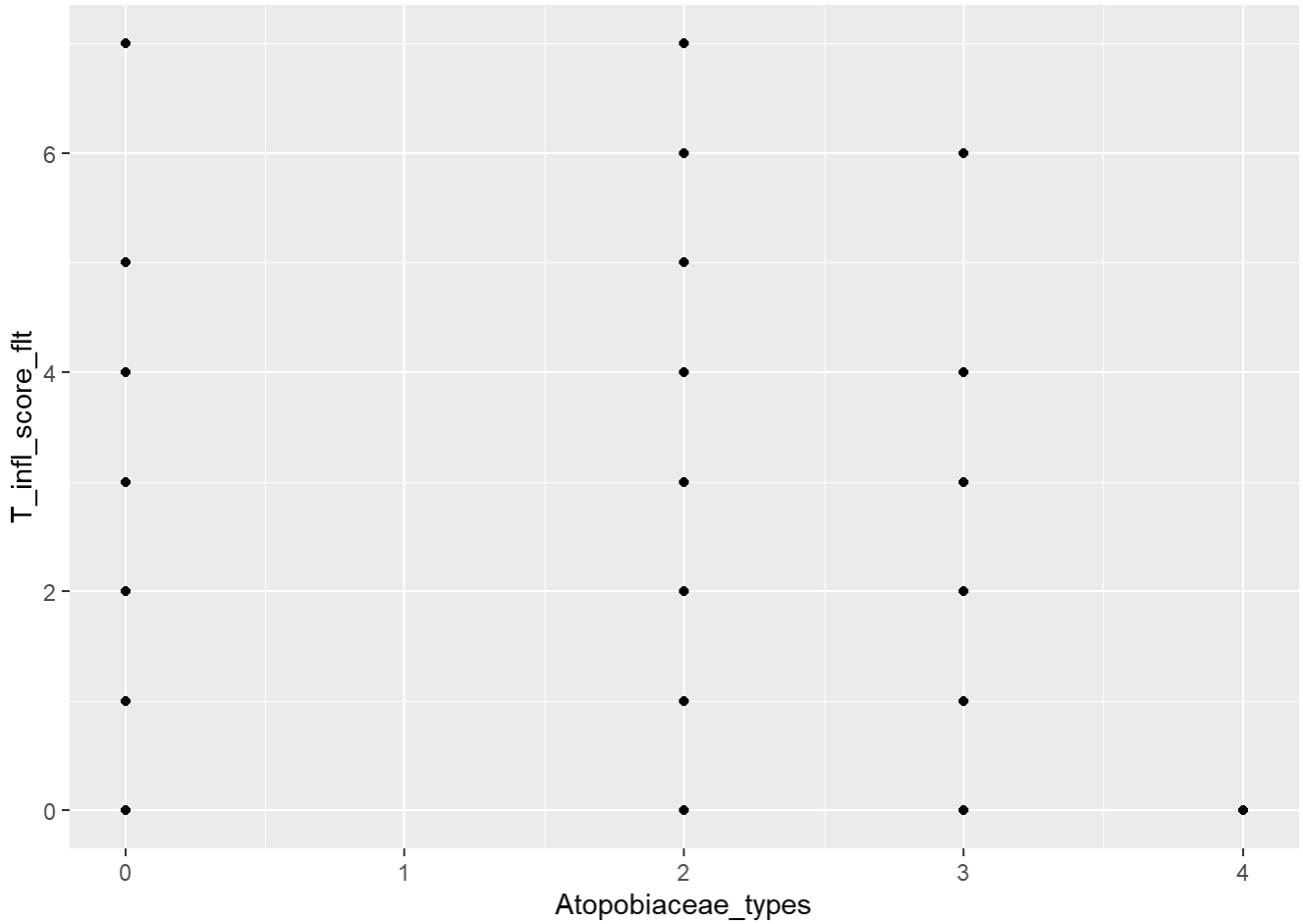
```
ggplot(Cer_AV_vag_meta, aes(x = Atopobiaceae_types, y = T_infl_score_flt)) +
  geom_point() +
  stat_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 1032192 rows containing non-finite values (stat_smooth).
```

```
## Warning: Computation failed in `stat_smooth()`:
## x has insufficient unique values to support 10 knots: reduce k.
```

```
## Warning: Removed 1032192 rows containing missing values (geom_point).
```



Since the data is in groups there is no correlation observed, let's try with more true continous variables… The null hypothesis is not rejected.

New Statistical Question: Does the number of BMI affect the genital inflammatory score?

Null hypothesis: The number of BMI does not affect genital inflammatory score Alternate hypothesis: the number of BMI does affect the genital inflammatory score

```
model <- lm(T_infl_score_flt ~ F_pcov_BMI, data = Cer_AV_vag_meta)
summary(model)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  3.35602759 0.76199269  4.404278 2.783624e-05
## F_pcov_BMI  -0.05263448 0.02476899 -2.125016 3.618035e-02
```

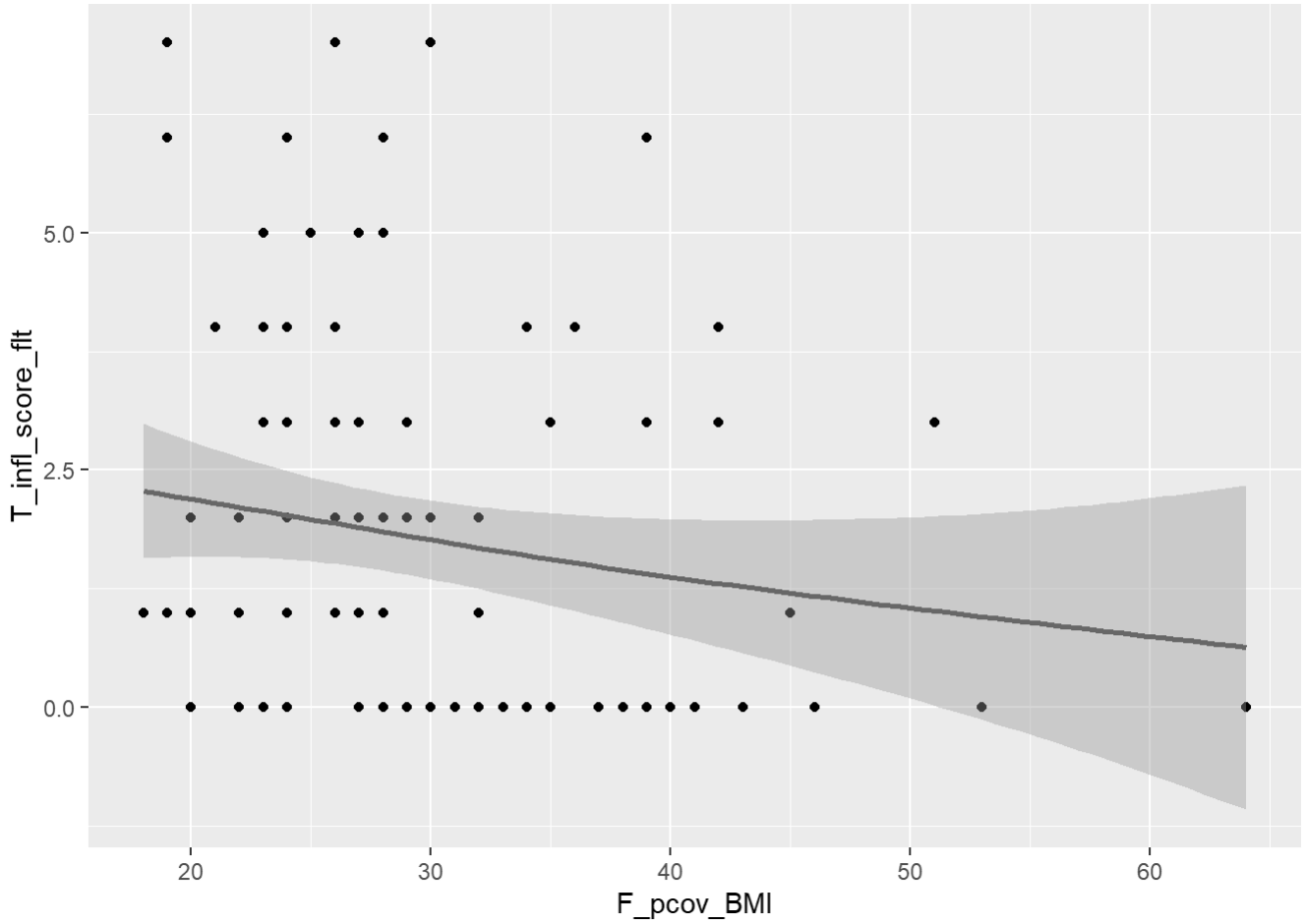
genital inflammation = 3.356 - 0.052*BMI

```
ggplot(Cer_AV_vag_meta, aes(x = F_pcov_BMI
, y = T_infl_score_flt)) +
  geom_point() +
  stat_smooth()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 1032193 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1032193 rows containing missing values (geom_point).
```



In this dataset there is a weak negative correlation observed with genital inflammation score. Meaning women with higher BMI tend to have lower genital inflammation which is counter-intuitive since obesity is associated with systemic inflammation. That said the inflammation level is a score an not based on a true continous variable such as inflammatory cytokine levels. The null hypothesis is rejected.