chp5

Vasco Brazão

15/02/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(here)
```

```
## here() starts at C:/Users/admin/Documents/statistical-rethinking
```

```
library(brms)
```

```
## Loading required package: Rcpp

## Loading 'brms' package (version 2.14.4). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').

##
## Attaching package: 'brms'

## The following object is masked from 'package:stats':
##
##     ar
```

5E1

(4) would be the standard way to write a multiple linear regression. I Suppose (2) could be valid? If we force the intercept to be 0. (3) seems plausible, but from the lack of an index on beta I would think you're forcing the beta for x to be equal to -1 * the beta for z, which is.. strange?

Van Bussel agrees https://github.com/castels/StatisticalRethinking/blob/master/Chapter%205/VanBussel_Chapter5_Questions.pdf

5E2

mu_latitude_i = alpha + beta_adiv * adiv_i + beta_pdiv * pdiv_i

5E3

$$time_i \sim \text{Normal}(\mu, \sigma)$$

$$\mu_i = \alpha + \beta_f f_i + \beta_s s_i$$

Both slope parameters should be positive.

Van Bussel agrees!

But I still can't make a stupid latex document. One day.

5E4

1, 3, 4 would be my guesses.

Van Bussel disagrees - 4 is not correct. But I still think it works?

And a latex document thing was created! I cannot believe my eyes. What fresh hell awaits me now? We shall see.. we shall see.

5M1

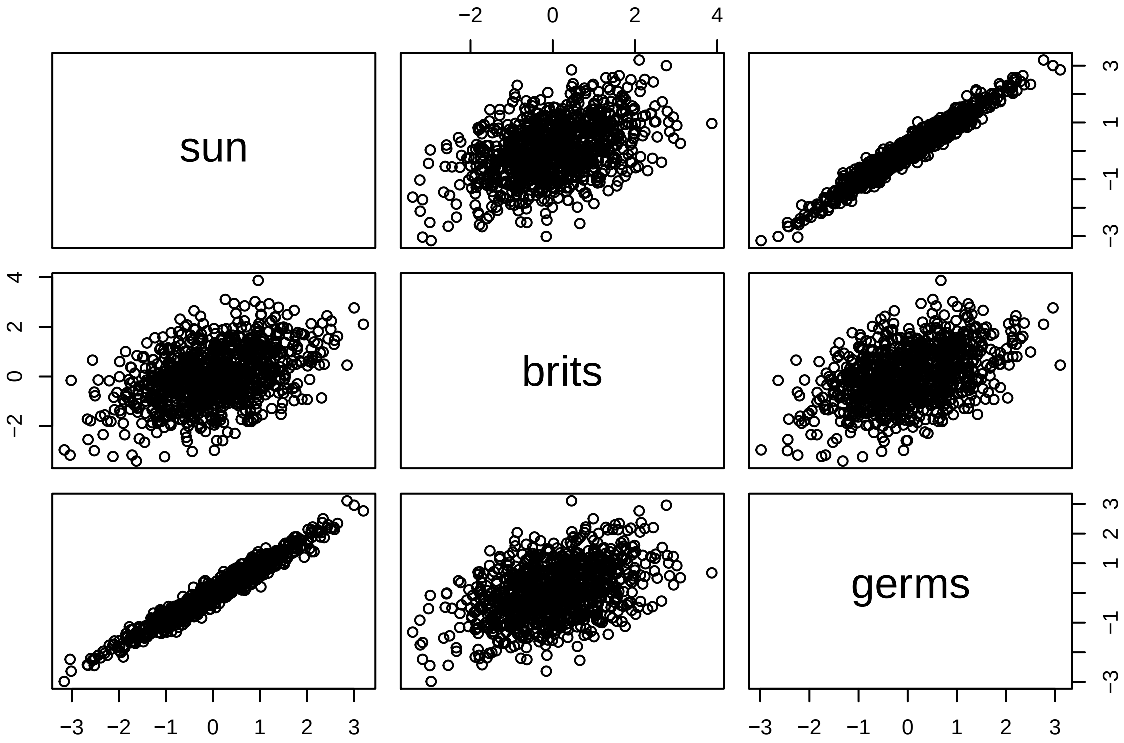
Inventing a spurious correlation.

Let's say that both British and German tourists like going to sunny places. For a given year, they base their decision of whether to go Portugal for vacation on the number of sunlight days of the previous year. Germans respond more strongly to this parameter because they now have an easier time going to Portugal than Brits, and they also stray less from the line because they are German.

```
n <- 1000

df <- tibble(
  sun = rnorm(n, 0, 1),
  brits = 0.5*sun + rnorm(n, 0, 1),
  germs = 0.9*sun + rnorm(n, 0, 0.2)
)
```

```
pairs(df)
```



There appears to be a positive relationship betweet amount of brits and amount of germans.

```
m1 <- lm(df$brits ~ df$germs)
m2 <- lm(df$brits ~ df$germs + df$sun)

summary(m1)
```

```
##
## Call:
## lm(formula = df$brits ~ df$germs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9256 -0.7035 -0.0363  0.7188  3.5215
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.008299   0.031633  -0.262    0.793
## df$germs     0.526425   0.034292  15.351   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9993 on 998 degrees of freedom
## Multiple R-squared:  0.191,  Adjusted R-squared:  0.1902
## F-statistic: 235.7 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = df$brits ~ df$germs + df$sun)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9551 -0.7106 -0.0299  0.7101  3.4592
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.009827   0.031610  -0.311   0.7560
## df$germs     0.253699   0.156998   1.616   0.1064
## df$sun       0.257862   0.144865   1.780   0.0754 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9982 on 997 degrees of freedom
## Multiple R-squared:  0.1936, Adjusted R-squared:  0.192
## F-statistic: 119.7 on 2 and 997 DF,  p-value: < 2.2e-16
```

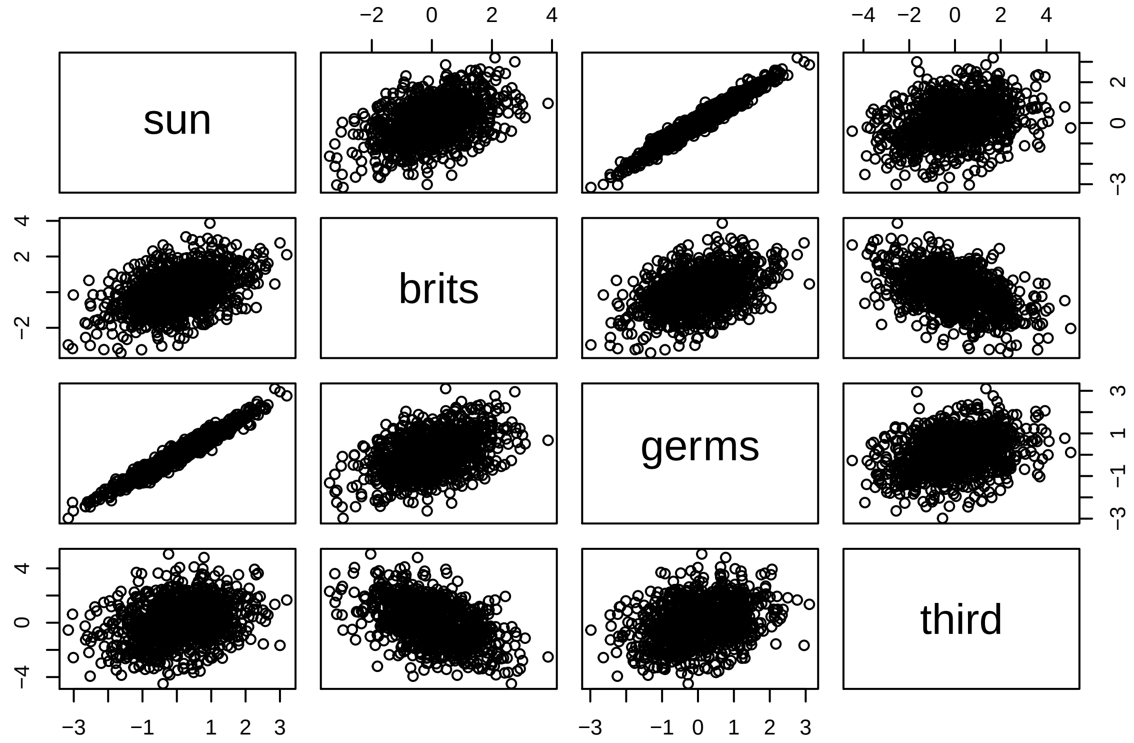
Et voila. when sun is in the model, knowing how many germans went to Portugal does not really tell us much more about how many brits went.

5M2

Now imagine a third country. Those citizens go to Portugal in higher numbers when they know their German friends will be going, but they prefer to avoid Brits on the beaches, for whatever reason.

```
df <- df %>%
  mutate(
    third = rnorm(n, mean = germs - brits)
  )

pairs(df)
```



```
lm(third ~ germs, data = df)
```

```
##
## Call:
## lm(formula = third ~ germs, data = df)
##
## Coefficients:
## (Intercept)        germs
##   -0.009561     0.491150
```

```
lm(third ~ brits, data = df)
```

```
##
## Call:
## lm(formula = third ~ brits, data = df)
##
## Coefficients:
## (Intercept)        brits
##      0.0204      -0.6432
```

```
lm(third ~ germs + brits, data = df)
```

```
##
## Call:
```

```
## lm(formula = third ~ germs + brits, data = df)
##
## Coefficients:
## (Intercept)        germs        brits
##    -0.01799      1.02569     -1.01542
```

Indeed! When both are included in the model, the coefficients are much greater. When only one is included, its relationship is masked.

5M3

If, in one year, a person marries, divorces, and marries again, they will have added two counts of marriage for that one year – marriage rate is affected by remarriages. Thus, in a society where divorce is accepted but being married is highly desirable, we could reasonably assume a casual link wherein the divorce rate influences the marriage rate. In years with (say) 0 divorces, marriages would mostly be first marriages. In years with a very high number of divorces, we might see that same baseline of new marriages PLUS a surge of remarriages driving up the marriage rate.

You could do a simple linear regression, though that would not inform you about the direction of causality. Alternatively, if you have reason to believe that the time between divorce and remarriage is about a year, you should see an effect of divorce at year 0 on marriage at year 1, but not so much the other way around.

5M4

Found the LDS data on wikipedia.

```r
data(WaffleDivorce, package = "rethinking")

d <- WaffleDivorce %>% as_tibble()
lds <- readxl::read_xlsx(path = here("chp5/ldsdata.xlsx"),
                         col_names = TRUE)

d2 <- full_join(d, lds, by = c("Location" = "State")) %>%
  select(Loc, MedianAgeMarriage, Marriage, Divorce, LDS) %>%
  filter(!is.na(Loc)) %>%
  mutate(
    across(.cols = MedianAgeMarriage:LDS,
           .fns = ~ rethinking::standardize(.))
  )
```

Now it's all standardized and in `d2`.

```r
lm(Divorce ~ MedianAgeMarriage + Marriage + LDS, data = d2) %>% summary
```

```
##
## Call:
## lm(formula = Divorce ~ MedianAgeMarriage + Marriage + LDS, data = d2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05378 -0.47978 -0.05498  0.53054  1.95093
##
```

```
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.659e-16  1.084e-01   0.000   1.0000
## MedianAgeMarriage -7.555e-01  1.604e-01  -4.709 2.32e-05 ***
## Marriage           6.338e-03  1.649e-01   0.038   0.9695
## LDS               -3.446e-01  1.297e-01  -2.658   0.0108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7667 on 46 degrees of freedom
## Multiple R-squared:  0.4482, Adjusted R-squared:  0.4122
## F-statistic: 12.45 on 3 and 46 DF,  p-value: 4.347e-06
```

```
b5m4 <-
   brm(data = d2,
       family = gaussian,
       Divorce ~ 1 + MedianAgeMarriage + Marriage + LDS,
       prior = c(prior(normal(0, 0.2), class = Intercept),
                 prior(normal(0, 0.5), class = b),
                 prior(exponential(1), class = sigma)),
       iter = 2000, warmup = 1000, chains = 4, cores = 4,
       seed = 5,
       sample_prior = T,
       file = "b5m4")

print(b5m4)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: Divorce ~ 1 + MedianAgeMarriage + Marriage + LDS
##    Data: d2 (Number of observations: 50)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##                   Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept            -0.00      0.10    -0.20     0.20 1.00     4078     2604
## MedianAgeMarriage    -0.68      0.15    -0.97    -0.38 1.00     3129     2736
## Marriage              0.04      0.16    -0.25     0.35 1.00     2856     2894
## LDS                  -0.31      0.13    -0.57    -0.04 1.00     3312     2811
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.78      0.08     0.64     0.96 1.00     3418     2103
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

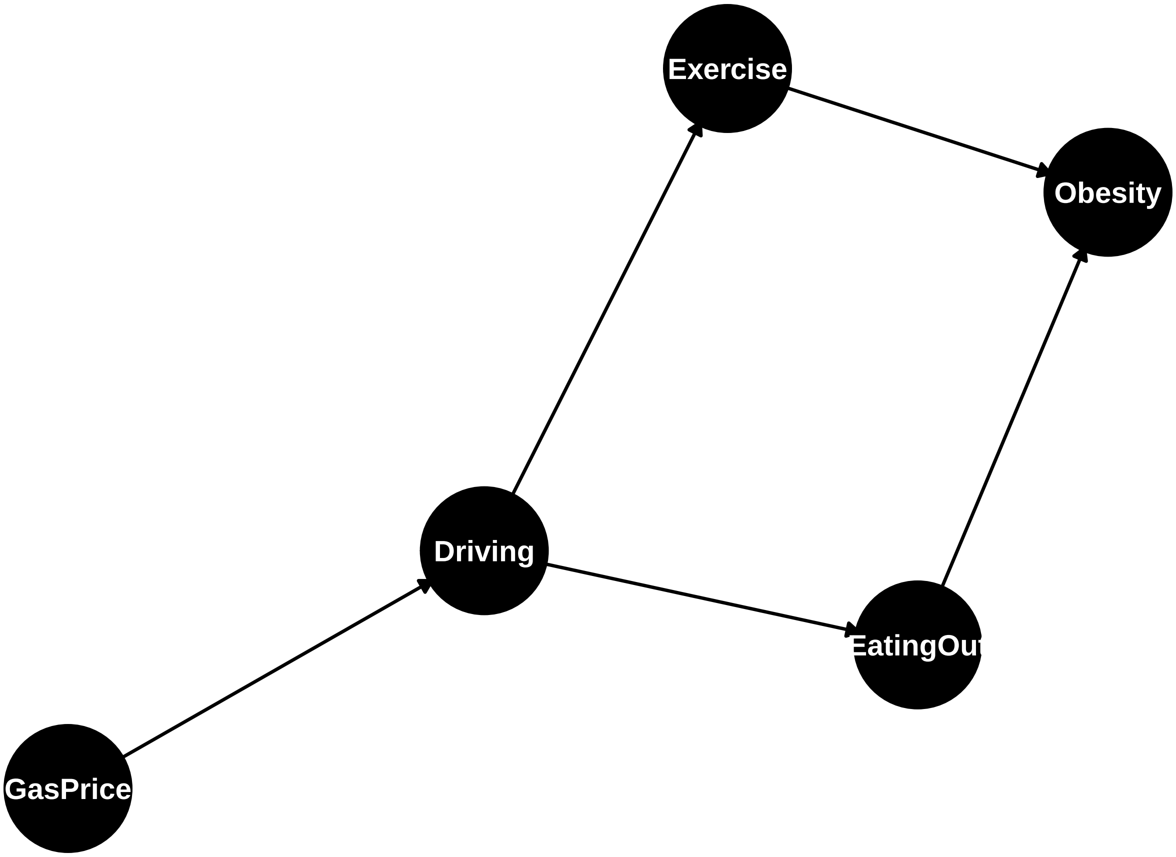
```
summary(b5m4)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: Divorce ~ 1 + MedianAgeMarriage + Marriage + LDS
##    Data: d2 (Number of observations: 50)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Population-Level Effects:
##                   Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept            -0.00      0.10    -0.20     0.20 1.00     4078     2604
## MedianAgeMarriage    -0.68      0.15    -0.97    -0.38 1.00     3129     2736
## Marriage              0.04      0.16    -0.25     0.35 1.00     2856     2894
## LDS                  -0.31      0.13    -0.57    -0.04 1.00     3312     2811
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.78      0.08     0.64     0.96 1.00     3418     2103
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

5M5

A DAG might be useful here.

```
ggdag::dagify(
  Obesity ~ Exercise + EatingOut,
  Exercise ~ Driving,
  EatingOut ~ Driving,
  Driving ~ GasPrice
) %>%
  ggdag::ggdag(node_size = 22) +
  theme_void()
```



How ugly! And it changes every time I run it. I embrace the chaos.

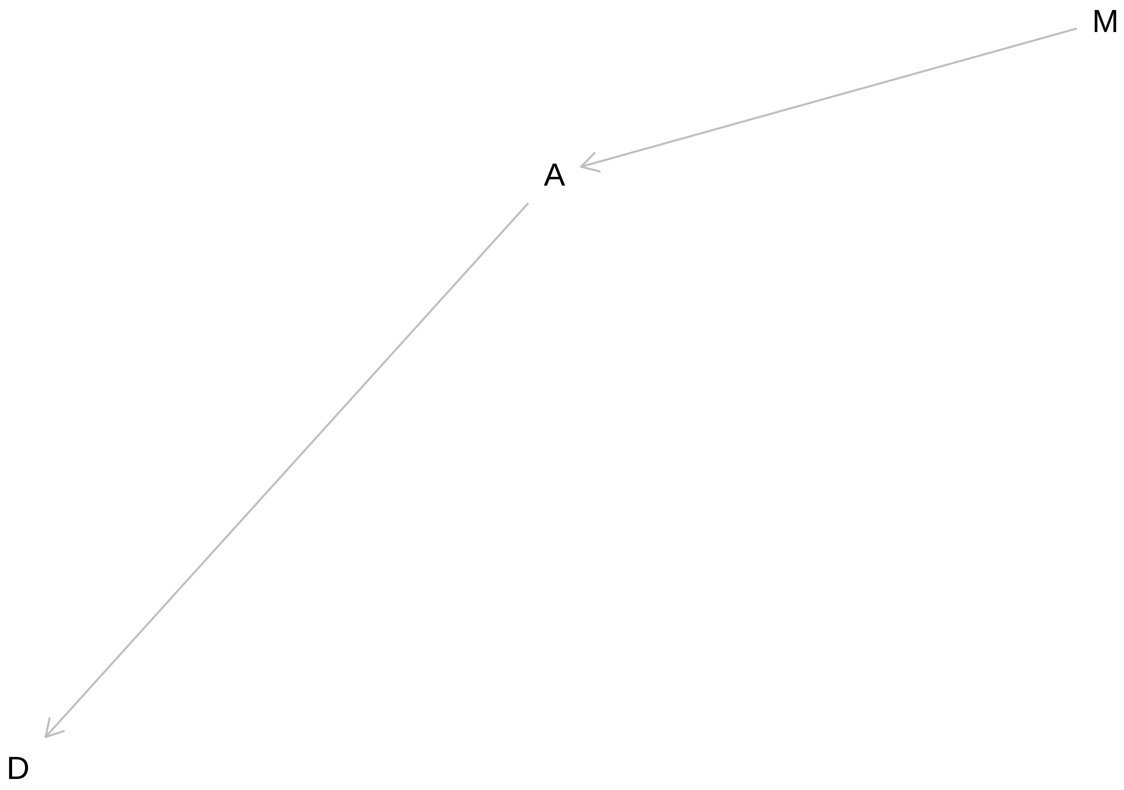
Ok, so…

- We could regress obesity on each of the predictors separately, just to confirm our intuitions.
- We could regress obesity on gas price and driving. We would expect that the coefficent for GasPrice would shrink, since Driving would be holding most of the information
- We could regress obesity on EatingOut and Exercise. The coefficients should shrink because they are collinear.
- We could regress obesity on EatingOut, Exercise, and Driving. The coefficient for Driving should shrink?

5H1

```
dag <- dagitty::dagitty(
  "dag{M -> A; A -> D}",
  layout = TRUE
)

plot(dag)
```



```
dagitty::impliedConditionalIndependencies(dag)
```

```
## D _||_ M | A
```

Without conditioning, all pairs of variables should be associated.

D is independent of M, conditioned on A. (A here is the mediator)

This is consistent with the pattern of results from the three models (m5.1, m5.2, and m5.3).