Problem set 20: Regression and correlation

Dmytro Shytikov

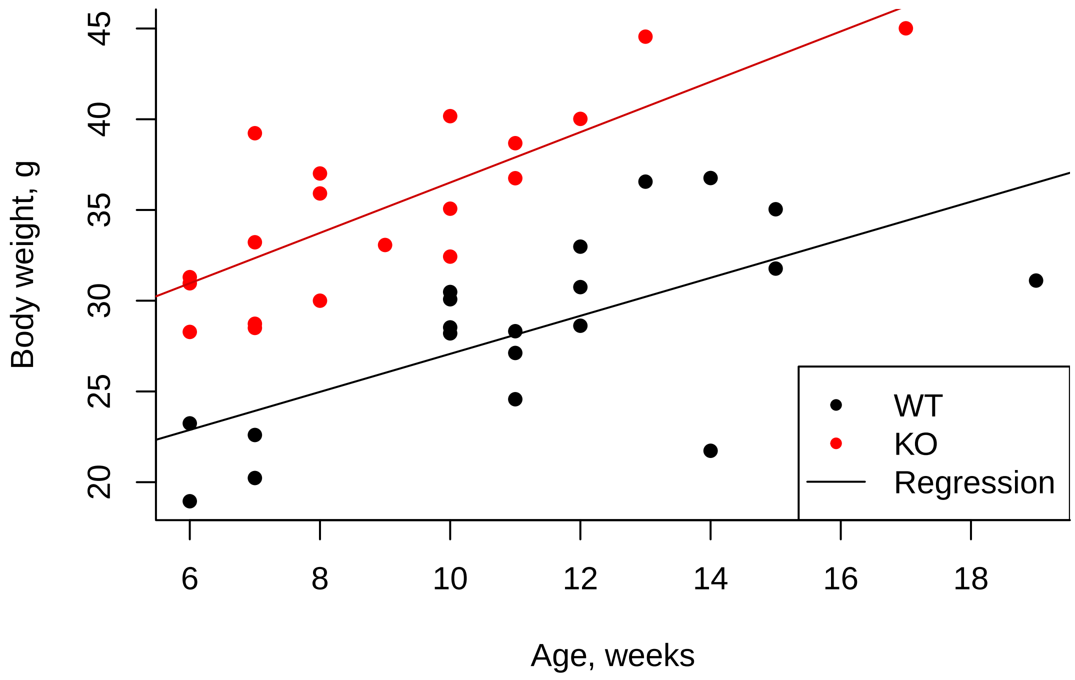
2024-03-19

Compare linear regression and correlation in two groups

Let's get the data into R:

```r
Mice_WT <- read.csv("WT.csv")
Mice_KO <- read.csv("KO.csv")
Mice_WT$Genotype <- "WT"
Mice_KO$Genotype <- "KO"
mice <- rbind(Mice_WT, Mice_KO) %>%
  mutate(Sex = as.factor(Sex),
         Genotype = as.factor(Genotype))
```

Let's see whether there is any difference in the regression slope between these two groups: WT and KO mice (Figure 1).



It may be interesting to compare the relationship between the age and body weight of WT and KO mice. There are several possible ways to do that. Except for including the genotype into the linear model as another explanatory variable (which you may do later as well), you may compare the regression and correlation coefficients (β and r, respectively) of these groups. After all, all these coefficients are estimates that can be compared in the hypothesis test. For example, in the z-test. You will need to do the following to compare the regression coefficients[1] (a simplified procedure):

$$z = \frac{\beta_{Group\ 1} - \beta_{Group\ 2}}{\sqrt{(SE_{beta_1}{}^2 - SE_{beta_2}{}^2)}}$$

```
model_wt <- summary(lm(Weight ~ Age, Mice_WT))
model_ko <- summary(lm(Weight ~ Age, Mice_KO))
# test assumptions on your own

model_wt
```

```

Call:
lm(formula = Weight ~ Age, data = Mice_WT)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5335 -1.8841  0.2795  2.7984  6.3443 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   16.594      3.241   5.120 7.16e-05 ***
Age            1.048      0.277   3.782  0.00136 ** 
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 3.974 on 18 degrees of freedom
Multiple R-squared:  0.4428,    Adjusted R-squared:  0.4119 
F-statistic: 14.31 on 1 and 18 DF,  p-value: 0.001364
```

```
model_ko
```

```

Call:
lm(formula = Weight ~ Age, data = Mice_KO)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0829 -2.2114  0.1691  1.4876  6.8811 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  22.6329     1.9740  11.465 1.05e-09 ***
Age           1.3880     0.2127   6.524 3.92e-06 ***
---
```

[1] Paternoster R, Brame R, Mazerolle P, Piquero A. Using the correct statistical test for the equality of regression coefficients. Criminology. 1998;36(4):859–66. Available from: http://dx.doi.org/10.1111/j.1745-9125.1998.tb01268.x

```
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 3.07 on 18 degrees of freedom
Multiple R-squared:  0.7028,    Adjusted R-squared:  0.6863
F-statistic: 42.57 on 1 and 18 DF,  p-value: 3.92e-06
```

```r
SE_b <- sqrt((model_wt$coefficients[2, 2])^2 + (model_ko$coefficients[2, 2])^2)
Z_age <- ((model_wt$coefficients[2, 1] - model_ko$coefficients[2, 1])/SE_b)
```

So, you can see that the difference between βs of both groups is -0.34 with the p-value of 0.33. Generally speaking, both types of mice gain weight at quite a similar rate and KO mice may have only a marginally steeper slope. Shame. Or… You may wish to add more variables into this model (try later)!

Correlation coefficients can be compared in a pretty straightforward way. Some additional R libraries may be handy.

Several explanatory variables at once

Theoretically, we have four possible explanatory variables. Let’s start from WT mice only and all four variables included in the model. Check the assumptions by yourself, and it is up to you to interpret whether it is possible to use this model, but compared to the original model ($Weight,\ g = Age \cdot \beta_{age} + intercept$), the multifactorial model does explain variance better: its RSE is lower and R^2 is higher.

```r
model_age <- lm(Weight ~ Age, data = Mice_WT)
model_multi <- lm(Weight ~ Age + Sex + Number + Tail, data = Mice_WT)

# The original univariate model
summary(model_age)
```

```
Call:
lm(formula = Weight ~ Age, data = Mice_WT)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5335 -1.8841  0.2795  2.7984  6.3443

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   16.594      3.241   5.120 7.16e-05 ***
Age            1.048      0.277   3.782  0.00136 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 3.974 on 18 degrees of freedom
Multiple R-squared:  0.4428,    Adjusted R-squared:  0.4119
F-statistic: 14.31 on 1 and 18 DF,  p-value: 0.001364
```

```r
# The model with all variables included
summary(model_multi)
```

```
Call:
lm(formula = Weight ~ Age + Sex + Number + Tail, data = Mice_WT)

Residuals:
   Min     1Q Median     3Q    Max
-5.706 -1.234 -0.529  1.923  4.311

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.0718    28.0093   1.038  0.31574
Age          -0.4015     0.7814  -0.514  0.61485
SexMale       5.3436     1.5929   3.355  0.00434 **
Number       -0.3309     0.3080  -1.074  0.29968
Tail          0.5320     1.8450   0.288  0.77701
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 3.002 on 15 degrees of freedom
Multiple R-squared:  0.735, Adjusted R-squared:  0.6643
F-statistic:  10.4 on 4 and 15 DF,  p-value: 0.0003078
```

But is this multifactorial model really good? Let's test what will happen with the unexplained variance if you add more factors and whether the observed drop in the unexplained variance is statistically significant. We will use the `anova(...)` function:

```r
anova(model_age, model_multi)
```

```
Analysis of Variance Table

Model 1: Weight ~ Age
Model 2: Weight ~ Age + Sex + Number + Tail
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     18 284.23
2     15 135.19  3    149.04 5.5123 0.009371 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

As you can see, the multivariate model does reduce RSS (unexplained variance) and these two models differ quite significantly. You can read the details about this procedure elsewhere[2].

However, you may notice that very few regression coefficients in the more complete model are significantly different from 0 (the respective p-values are often very high). So, it does explain variance better, but very few regression coefficients seem to differ from 0. How comes? Check the assumptions! Probably, the problem is somewhere there.

Now, you may refine your model and exclude some parts of the equation (for example, remove problematic or useless explanatory variables). You can do it "by hand" first (and do not forget to test assumptions *before* you make any conclusions!). Or you may do a stepwise model selection. Again, be mindful when you include different variables. Also, remember that more does not always mean better. If your model includes too many

[2] 16.5: The F test as a model comparison [Internet]. Statistics LibreTexts. Libretexts; 2019 [cited 2024 Feb 27]. Available from: https://stats.libretexts.org/Bookshelves/Applied_Statistics/Learning_Statistics_with_R_-_A_tutorial_for_Psychology_Students_and_other_Beginners_(Navarro)/16%3A_Factorial_ANOVA/16.05%3A_The_____F_____test_as_a_model_comparison

parameters (more or almost the same number as observations), it will be also unreliable as it explains too much variance leaving too little for the unexplained variance!

Stepwise regression can be done by this command:

```r
step(object = model_1, # The model from which you are going to start
     direction = "backward",  # Choose "forward", "backward", or "both"
     scope = formula(another_model), # The model to which you are aiming
     ...) # Check the other possible arguments on your own
```

This code will add more variables to your model, remove variables, or do both depending on the `direction` argument. If the new model with (or without) a certain term is better than the previous one (its Akaike Information Criterion (AIC)[3] decreases), it will add (or delete) it to the previous model and try to iterate the cycle again. If the addition (or removal) of the model term increases the AIC score, the iterations are stopped and the last formulated model will be preserved. You may try something like this:

```r
step(
  object = lm(Weight ~ 1, data = mice), # The model with no predictors
  direction = "forward",
  scope = formula(lm(Weight ~ ., mice[,-1])) # The model with all variables
  # Except for the ID number
)
```

```
Start:  AIC=146.47
Weight ~ 1

           Df Sum of Sq    RSS    AIC
+ Genotype  1    400.25 1080.8 135.87
+ Age       1    245.98 1235.1 141.20
+ Tail      1    211.80 1269.3 142.29
+ Sex       1    193.51 1287.6 142.87
<none>                  1481.1 146.47

Step:  AIC=135.86
Weight ~ Genotype

       Df Sum of Sq     RSS    AIC
+ Age   1    615.03  465.81 104.20
+ Tail  1    520.04  560.80 111.62
+ Sex   1    193.51  887.34 129.97
<none>              1080.84 135.87

Step:  AIC=104.2
Weight ~ Genotype + Age

       Df Sum of Sq    RSS    AIC
+ Sex   1    26.215 439.60 103.88
<none>              465.81 104.20
+ Tail  1    19.269 446.54 104.51

Step:  AIC=103.88
```

[3] Bevans R. Akaike information criterion [Internet]. Scribbr. 2020 [cited 2024 Feb 27]. Available from: https://www.scribbr.com/statistics/akaike-information-criterion/

```
Weight ~ Genotype + Age + Sex

       Df Sum of Sq    RSS    AIC
<none>              439.60 103.88
+ Tail  1    16.143 423.46 104.38
```

```
Call:
lm(formula = Weight ~ Genotype + Age + Sex, data = mice)

Coefficients:
(Intercept)   GenotypeWT          Age      SexMale
     23.786       -9.180        1.119        1.826
```

As you can see, the original model without any variabes and with only the intercept term (`lm(Weight ~ 1, data)`) has the AIC score = 146.7. After we add `Genotype` to the model, the AIC score reduces to 135.9. After adding `Age`, it further drops to 104.2, etc. Thus, we see how the model evolves. At the same time, `Tail` does not improve the model, so it is not included. And the ID number was not included because it is pointless (discussed earlier). The final lines of the output suggest the finalized multiple regression model that may be good enough to explain the weight of mice. Now, you can pick it and analyze it further. Or find an even more optimized model!

You may wish to check the correlation between the dependent and explanatory variables in this model. Don't be tempted to run `cor.test()` at once! It will give you heavily biased results. Use partial correlation instead. Read about it elsewhere.

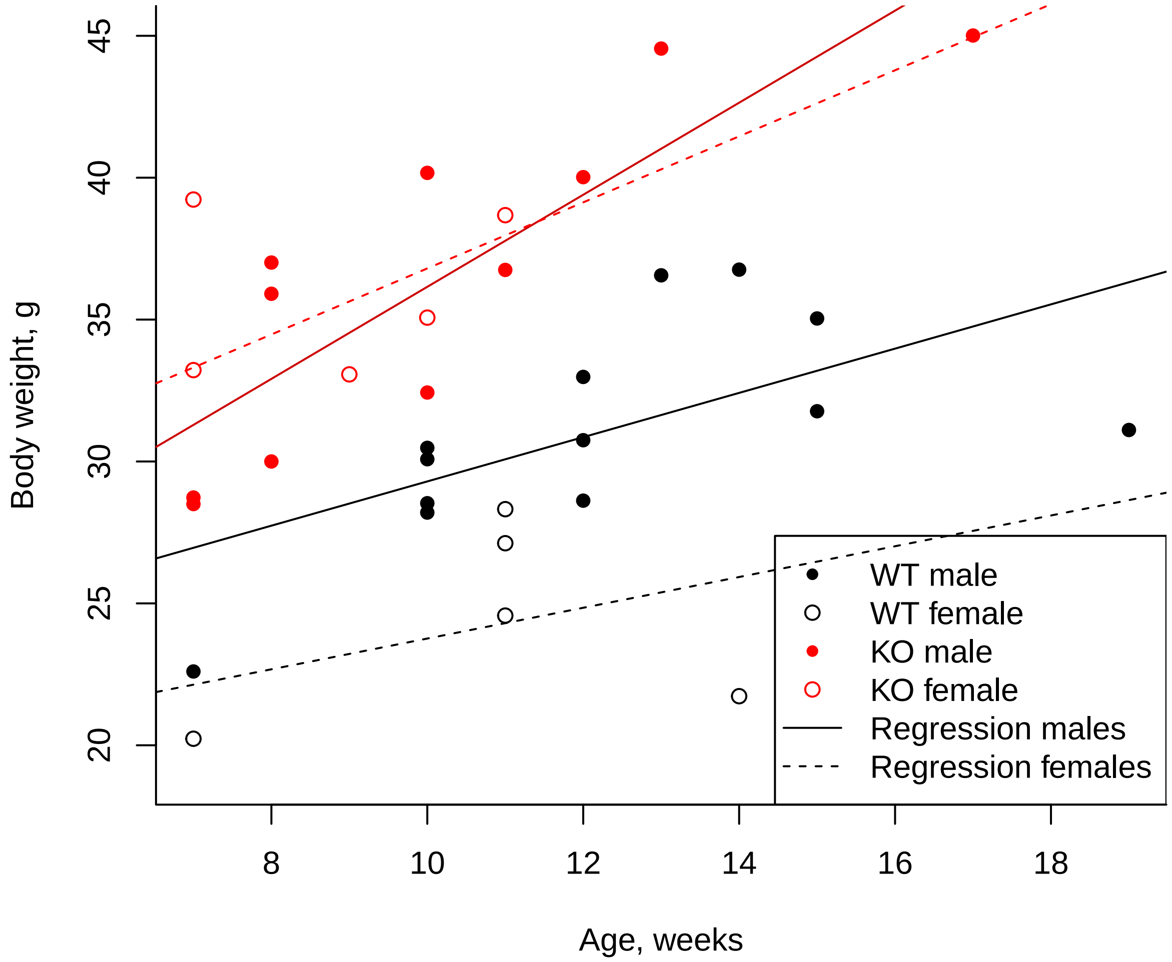
One question may arise: how do we work with categorical variables such as Sex or Genotype? There are only two possible values here. In this case, your equation may look like this: $Weight,\ g = 0 \cdot \beta_{Sex} + intercept = intercept$ for males and $Weight,\ g = 1 \cdot \beta_{Sex} + intercept$ for females (or vice versa). A more interesting situation will happen if there are several levels of the categorical variable, but it is beyond the scope of the seminar.

Especially interesting may be inclusion not only the main effects of certain (categorical) variables, but also the product of their interactions (remember this term from the ANOVA-related tutorial?). But by doing so, we will enter another field of analysis, analysis of covariance (ANCOVA), which is also beyond the scope of this tutorial. And again, beware: you must choose your model wisely, do not include terms or factors just because you can!

For the sake of interest, compare these models:

```r
model_3 <- lm(formula = Weight ~ Genotype + Age + Sex, data = mice)
model_3int <- lm(formula = Weight ~ Age + Genotype * Sex, data = mice)
model_3full <- lm(formula = Weight ~ Age * Genotype * Sex, data = mice)
```

Finally, you may wish to plot your results (Figure 2):



Originally created by Dmytro Shytikov in 2024