Problem set 20: Regression and correlation

Dmytro Shytikov

2024-03-19

Compare linear regression and correlation in two groups

You have identified several variables that can explain differences in the body weight of your WT mice. For example, age. However, exploring WT mice is not as interesting as comparing them to mice of another genotype. Let's say, mice that will spontaneously develop obesity. Load `KO.csv` and analyze it as well. Does the data look different compared with the WT mice you analyzed before?

Probably, it would be interesting to compare how 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):

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

This information you can get from your `summary(lm(...))` table. However, this formula is somewhat simplified and may not be precise enough. You may look for better formulas elsewhere (for example, you may check Andrade and Estévez-Pérez, 2014[2]).

To compare the correlation coefficients of two groups, you would need to convert them to z-values and perform a z-test on them just as above (Fisher Z-transformation[3,4]):

$$r\ to\ z\ transformation : z_r = \frac{1}{2} ln(\frac{1+r}{1-r})$$

$$SE_{difference\ of\ z_r} = \sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}$$

$$z = \frac{z_{r_1} - z_{r_2}}{SE_{difference\ of\ z_r}}$$

Several explanatory variables at once

You have discussed the link between several explanatory variables (`Age`, for example) and the body weight of the mice during this week's tutorial. But what if you formulate a model that will include several explanatory variables at once? Something like this:

[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

[2] Andrade JM, Estévez-Pérez MG. Statistical comparison of the slopes of two regression lines: A tutorial. Anal Chim Acta. 2014;838:1–12. Available from: http://dx.doi.org/10.1016/j.aca.2014.04.057

[3] Fisher. RA. Statistical Methods for Research Workers. Edinburgh, Scotland,: Oliver and Boyd; 1925.

[4] Or elsewhere

```r
# A model without interaction
lm(dependent_var ~ explanatory_var_1 + explanatory_var_2, data = data_frame)

# A model with interaction
lm(dependent_var ~ explanatory_var_1 * explanatory_var_2, data = data_frame)
```

Especially given you have several more variables that could become explanatory (which one would you choose?). By doing so, you will start multiple regression analysis, which is one of the exciting fields in data analysis! But beware: the requirements for your analysis become even stricter. And it is easy to get lost in different models! But let's start by specifying the model. Which variables to include? Theoretically, we have four possible explanatory variables! Let's start with all four variables included and our WT mice!

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

First of all, assumptions. You must make sure about:

- Independence of observations;
- Linear relationship between variables;
- Homoscedacity of residuals;
- Normal distribution of residuals of the multivariate model;
- **No correlation between the explanatory variables (multicollinearity)**.

You already know how to run the first four assumptions (you did that in the tutorial). As for multicollinearity, the model becomes unreliable if two or more predictor variables are highly correlated to each other. You can test multicollinearity by determining the variance inflation factor (VIF). It measures the correlation between the explanatory variables within your model.

You can do it in R by using the `vif()` function of the `car` package:

```r
library(car)
vif(your_model)
```

The value for VIF starts at 1 and has no upper limit. A general rule of thumb for interpreting VIFs is as follows[5,6]:

- A value of 1 indicates there is no correlation between a given predictor variable and any other predictor variables in the model.
- A value between 1 and 5 indicates a moderate correlation between a given predictor variable and other predictor variables in the model, but this is often not severe enough to require attention.
- A value greater than 5 indicates a potentially severe correlation between a given predictor variable and other predictor variables in the model. In this case, the coefficient estimates and p-values in the regression output are likely unreliable.

Check what your model shows. Is it something we should care about? Let's check what this model will give us! (**Normally, you *must* test the assumptions first, and treat your data if some of them are violated. This time, for educational purposes, we will do the analysis differently.**) Let's investigate it and compare it with the one that included only one variable first! For example:

[5] Zach. The five assumptions of multiple linear regression [Internet]. Statology. 2021 [cited 2024 Feb 27]. Available from: https://www.statology.org/multiple-linear-regression-assumptions/

[6] Zach. How to calculate variance inflation factor (VIF) in R [Internet]. Statology. 2019 [cited 2024 Feb 27]. Available from: https://www.statology.org/variance-inflation-factor-r/

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

summary(model_multi)
summary(model_age)
```

Are the results different? Which model explains the overall variance better? Which one would you choose? Did you expect these results? How would you interpret these results?

So, you have two different models. But which one would you prefer? You can compare them by using the `anova(...)` function:

```
anova(model_1, model_2, ..., model_n)
```

The `anova()` function will take the model objects as arguments, and return an ANOVA testing whether the more complex model is significantly better at capturing the data than the simpler model. Basically, it will check whether the variance explained by a more complex model will be higher than the unexplained error variance. If the resulting p-value is sufficiently low (usually less than 0.05), we conclude that the more complex model is significantly better than the simpler model, and thus favor the more complex model. If the p-value is not sufficiently low (usually greater than 0.05), we should favor the simpler model[7].

So, does the multivariate model make sense? Does it explain variance better than the univariate one? Probably, the multivariate model explains variance better, but it is still not good enough. You may wish to refine your model and exclude some parts of the equation. Try to do it by hand first (and do not forget to test assumptions *before* you make any conclusions!).

Too many models to test

This is a very simple data set, but it already gets annoying, especially if you include `Genotype`. How to choose the equation that is good enough? By just typing all the possible combinations? Luckily, you do not have to do it by hand.

You can run the stepwise model selection:

```
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 either add more variables to your model or reduce them, or do both. If the new model is better than the previous one, it will stop. Find a multivariate model that will describe your variables the best. You may also wish to assume interactions between some explanatory variables. Remember: you should be thoughtful when you choose the explanatory variables and when you formulate your model. Parts of the equation should be possible to interpret!

Originally created by Dmytro Shytikov in 2024

[7] Phillips N. YaRrr! The Pirate's Guide to R. APS observer [Internet]. 2017 [cited 2024 Feb 27];30. Available from: https://bookdown.org/ndphillips/YaRrr/comparing-regression-models-with-anova.html