**Lab 5: Correlation and Regression**

**5.1. Correlation**

The correlation coefficient gives the amount of covariance between a pair of continuous variables relative to the total amount of variance in the variables. We can illustrate this by using **rnorm()** to simulate variation in two traits (e.g. basal metabolism and sweating) which are genetically correlated due to sharing some of their underlying genes (some genes have effects on both traits). Both traits also vary independently from each other because some genes are private and affect one but not the other trait, and, because there is also environmental variance (measurement error) that affect estimates of both traits independently. Let’s create a dataset of 1000 study subjects and measure them for trait A and trait B. We measure the phenotypes of A and B, but the phenotypic variation among study subjects come from variation in the genes that are shared by the two traits, variation in the private genes for each trait, and the environmental variance in each trait.

**shared <- rnorm(1000, 0, 20)**

**private.A <- rnorm(1000, 0, 20)**

**private.B <- rnorm(1000, 0, 20)**

**env.A <- rnorm(1000, 0, 10)**

**env.B <- rnorm(1000, 0, 10)**

**A <- shared + private.A + env.A**

**B <- shared + private.B + env.B**

Try plotting the variables against each other (**plot(A, B)**) to see how that looks.

**5.1a *Can you calculate the sample correlations coefficient between A and B (i.e. your estimate of rho from your sample) using the sd() / var() and cov() functions in R?***

**5.1b Can you calculate the t*rue population correlation coefficient between A and B (rho) using the information above and the notes from lecture 5? (hint: the covariance between a variable and itself is equal to its variance: VAR(A) = COV(A,A))***

*Does your result make sense? Check your result using the* ***cor()*** *function. You can also test for significance using the* ***cor.test()****. Make sure you understand the output*.

**5.2. Simple Linear Regression**

Read in the dataset “**AmericanCrime.txt**” available in the folder of Lab 5. This data contains yearly overall rates of crime (per million) and violent crime (per 100.000) along with investment into the local police force and the percentage of adult residents having attended high school and college. For now, let’s only look at the two variables violent crime rate and police funding.

***5.2a Do you have an hypothesis you want to test with these variables? Should you use correlation or regression? Which is your dependent variable in case of regression?***

If you choose regression, perhaps you would like to investigate if the investment in police funding has any effect on violent crime rate, applying simple regression (including a single explanatory variable):

**simple.2a <- lm(violent.crime ~ police.funding , data = crime)**

Before looking at the results, make sure you inspect the model residuals. The inspection can be eased by setting up the graphical window so that it displays the four plots next to each other:

**par(mfrow=c(2,2))**

**plot(simple.2a)**

***5.2b Is the model OK? OR, do you need to deal with your data in order to fulfill the criteria for analysis? And, for regression analysis, that weird looking last panel down to the right actually is informative to look at – make sure you understand its general message. Talk to your fellow students and teacher.***

Run your model again and make sure that you have successfully dealt with any irregularities by plotting residuals: I choose transforming both variables using the logarithm with 10 as base – this looks at proportional relationship between the x and y variable, and log10 has a straightforward interpretation when looking at each variable separately, and gave decent residual distribution.

**simple.2b <- lm(log10(violent.crime) ~ log10(police.funding) , data = crime)**

Summarize your final model using **summary()**.

To help you interpret the result, you can also plot the relationship, e.g.:

**plot(log10(violent.crime) ~ log10(police.funding) , data = crime)**

You can add the regression slope to your plot with the command:

**abline(simple.2b)**

***5.2c Is the result what you predicted? Do you understand all the model output: can you interpret the regression coefficients? How is the t- and F-value calculated from the other values in the tables? Do the degrees of freedom make sense? What does the R2 value mean?***

The adjusted R2 value is a bit peculiar. The adjustment comes from a penalty paid by introducing more explanatory variables in your model. This does not make any sense in this case where you only have one variable in the model, but in cases where there are more than one explanatory variable it usually does make sense, because introducing any random variable to your dataset will ALWAYS explain additional variance in your dependent variable, just by chance. So the adjusted R2 gives you a more objective estimate if the model really was improved by adding more variables. We can illustrate this by adding nonsense variables to our analysis and comparing models:

**Number.of.oranges <- rnorm(length(crime$violent.crime), 100, 10)**

**Number.of.apples <- rnorm(length(crime$violent.crime), 100, 10)**

**multi.2c <- lm(log10(violent.crime) ~ log10(police.funding) + Number.of.oranges + Number.of.apples , data = crime)**

Look at the R2 versus the adjusted R2 in the models. You can also compare the models by:

**anova(simple.2b, multi.2c)**

The P-value from this comparison tells you whether including apples and oranges explained significantly more variation in crime (i.e. was the second more complex model significantly better than the first, less complex, model?).

You can also apply *robust regression* as an alternative to transforming your data or removing outliers. There are many implementations of robust regression in R. One basic implementation is available in the **MASS**-package. We can try to also run a robust regression on our original model:

**library(MASS)**

**simple.2d <- rlm(violent.crime ~ police.funding , data = crime)**

**summary(simple.2d)**

**library(car)**

**Anova(simple.2d)**

***5.2d Compare the P-value, t-value and coefficients between your original model, the model on transformed data, and the robust regression. Also compare the residual plots from each model. What did the robust regression do? Which model do you prefer, and why?***

**3. Multiple Linear Regression**

Let’s continue with “**AmericanCrime.txt**” but now look at all five variables. Inspect the data using the summary commands you learnt about previously (e.g. lab 1). You are interested in understanding and predicting crime.

***5.3a Which variables do you think should be considered as independent variables, and which do you consider to be dependent? Do you think there are causal relationships between your variables, or are they just correlated? Discuss with your bench-mates.***

Let’s look at the correlations between all variables in the dataset. When the data comes in matrix form, it is quite handy just to plot the whole dataset: **plot(crime)**. You can also try to load the **corrplot** package (you need to install it first if you haven’t) and run **corrplot(cor(crime))**. If you are not so much into graphics you can simply print the correlation matrix using **cor(crime)**. *How does it look – is it what you expected?*

Sometimes it can be hard to see the causal relationships between variables due to the collinearity among variables. For example, if we want to understand how the crime rate is related to police funding and education, we may need to analyze them simultaneously in multiple regression. Given that the two variables for crime (crime.rate and violent.crime) and education (high-school and college) are highly correlated, we can start by just analyzing one of each together with police funding. Let’s illustrate the benefits of multiple regression by first performing two separate simple regressions for each explanatory variable on the rate of violent crime, and then comparing the output (use **summary()**) to that of a multiple regression including both:

**uni.3a <- lm(log10(violent.crime) ~ log10(police.funding), data = crime)**

**uni.3b <- lm(log10(violent.crime) ~ perc.highschool, data = crime)**

**multi.3 <- lm(log10(violent.crime) ~ log10(police.funding) + perc.highschool, data = crime)**

***5.3b Do you notice anything in particular? What was the effect of including both variables simultaneously, and why did this happen?***

Visualizing relationships between more than 2 variables quickly becomes tricky and you will spend more time on using multivariate techniques to do so in the next couple of lectures and labs. For now, you can try one of the many R functions for plotting 3 variables; **scatterplot3d()**:

**library(scatterplot3d)**

**scatterplot3d(crime$police.funding, crime$perc.highschool, crime$violent.crime, pch=20)**

There are many options here. You can type ?scatterplot3d to find out more.

Now let’s try to explain the rate of violent crime by all four explanatory variables:

**multi.3b <- lm(log10(violent.crime) ~ log10(crime.rate) + log10(police.funding) + perc.highschool + perc.college, data = crime)**

***5.3c What happened? Discuss with your bench-mates.***

Collinearity between the two crime-variables (one as the response/dependent and one as an explanatory/independent variable) caused problems in that model. Collinearity can also cause problems when interpreting the effect of education when we put in both education-variables, even if we leave crime.rate out of the model:

**multi.3c <- lm(log(violent.crime) ~ log10(police.funding) + perc.highschool + perc.college, data = crime)**

When you have several highly correlated variables in your dataset, it can be hard for the model to figure out the true relationships. Here, we first included two measures of crime that naturally will be correlated, but we added one of the measures (overall rate of crime) as an explanatory variable. This resulted in police funding not explaining any additional variation in the rate of violent crime. Second, we put in two correlated measures of education. This resulted in the model hinting that college attendance might even be positively correlated to crime! To understand this problem better, you can compare the correlation matrix (**cor(crime)**) to the partial correlation matrix of the 5 variables. To get the partial correlation matrix we can install and load the **RcmdrMisc** package, and then run:

**partial.cor(crime)**

Naturally, we would like to model the two variables describing crime rate together as the response, and the other three as explanatory variables, with the two variables describing education possibly combined into one variable. Moreover, you may want some objective method that allows you to deal with these types of problems for data where the relationships between your set of variables (e.g. expressed genes) are hard to classify *a priori* based on any hypothesis. You will learn more about multivariate techniques that allow you to do this in the next couple of lectures and labs.