MATH 6359, Statistical Computing, Homework 3

Andrey Skripnikov

October 1, 2017

SUBMISSION GUIDELINES:

- **Bring the hard-copy of typed up solutions to the class on Tuesday, Oct 10.**
- **Keep it under 6 pages total (all included - text, code, plots, tables). DON'T (!) repeat the problem formulation, go straight to solution.**
- **For problems 1,3 and 4 follow the example format in terms of conciseness.**
- **For problem 2 - provide both code and your answers in plain English.**
- **Point total is 65 (100%), and on top of that one can get 5 extra credit points total.**

PROBLEM #1 - 25 points.

Write your own function performing a two-sample two-sided (don't do one-sided versions) proportion test via normal approximation (slide 12 in Part 5). As arguments the function has to take:

1. **a two-element vector containing numbers of successes for groups 1 and 2,**
2. **a two-element vector containing numbers of trials for groups 1 and 2,**
3. **a parameter that allows to specify if one wants to see the** *z-test* **or** *χ^2-test* **statistics**

The function has to output:

1. **TS, the calculated test-statistic value.**
2. **The TYPE of calculated test statistic (as inputted by the user, z- or χ^2)**
3. **p-value.**

4. **Vector containing sample proportions for groups 1 and 2.**

Having defined your function, first show a couple of basic calls, and a comparative call of *prop.test*() **function (see example). Afterwards, put it to work by performing the statistical practice of 1) generating your own simulation data; 2) checking if the two-sample test correctly identifies whether there is a difference between groups or not:**

1. **Consider two fixed binomial distributions -** $Bin(n, p_1)$ **and** $Bin(n, p_2)$

2. **Pick the following three combinations of values for** (p_1, p_2)**:**

   - p_1 **and** p_2 **are very different (e.g.** $p_1 = 0.82$ **and** $p_2 = 0.26$**)**
   - p_1 **and** p_2 **are slightly different (e.g.** $p_1 = 0.43$**,** $p_2 = 0.53$**)**
   - p_1 **and** p_2 **are exactly the same (e.g.** $p_1 = p_2 = 0.3$**)**

3. **For each of three combinations** (p_1, p_2)**:**

   - **Consider** $n = 10^2, 10^4, 10^6$ **and generate of random values from** $Bin(n, p_1)$ **and** $Bin(n, p_2)$ **for each** n **value**
   - **Calculate the numbers** x_1 **and** x_2 **of successes observed for each group**
   - **Run your two-sample testing function on the observed successes** (x_1, x_2) **and numbers of trials** (n, n) **to calculate the p-values**

Summarize the outputted p-values in a table. Do the results of the tests agree with the true underlying distributions (from which the data was generated)? What do you mostly witness for small samples ($n = 10^2$) **as opposed to big samples** ($n = 10^6$)**?**

EXAMPLE:

```
> two.samp.prop <- function(succ,n,stat.type="z"){
 ... That's all you. ....
}

## First show me a couple of test calls.
> two.samp.prop(c(23,25),c(40,50))  # See the output layout that I expect.
$TS                                 # Z-test statistic value.
[1] 0.7086834
$stat.type                          # Type of statistic ("z" in that case)
[1] "z"
$p.val                              # Self-explanatory.
[1] 0.478521
```

```
$sample.est                          # c(p1.hat,p2.hat)
[1] 0.575 0.500
>
> two.samp.prop(c(23,25),c(40,50),stat.type = "chisq") # Specify that you want chi-squared.
$TS                                                     # Chi-sq statistic value.
[1] 0.5022321
$stat.type                                              # Stat type - "chisq".
[1] "chisq"
$p.val
[1] 0.478521
$sample.est
[1] 0.575 0.500
>
> prop.test(c(23,25),c(40,50),corr=F)   # prop.test() output should match the prev. output.
2-sample test for equality of proportions without continuity correction
data:  c(23, 25) out of c(40, 50)
X-squared = 0.50223, df = 1, p-value = 0.4785 # Should be the same as your TS, df and p.val
alternative hypothesis: two.sided
95 percent confidence interval:                 # You are not required to calculate those.
 -0.1315822  0.2815822
sample estimates:
prop 1 prop 2                                   # Should be same as your sample.est.
 0.575  0.500
>
>
> p.set.1 <- c(0.82,0.26)
> p.set.2 <- c(0.43,0.52)
> p.set.3 <- c(0.3,0.3)
> for (n in c(10^2,10^4,10^6)){
  # Here separately for each p.set you:
  #   - Generate two random numbers of successes from respective binomial distributions
  #     with n trials (one number per group)
  #   - Feed the resulting generated numbers of successes
  #     alongside the numbers of trials to your function.
  #   - Print the p-values.
}
```

Table (please comment on your table in your work):

Table of p-values	$n = 10^2$	$n = 10^4$	$n = 10^6$
$p_1 = 0.82, p_2 = 0.26$	3.370715e-11	0	0
$p_1 = 0.43, p_2 = 0.52$	0.05966605	0	0
$p_1 = 0.3, p_2 = 0.3$	0.06375417	0.2203307	0.3876499

PROBLEM #2 - 10 points.

Two drugs for the treatment of peptic ulcer were compared. The results were as follows:

	Healed	Not Healed	Total
Pirenzepine	23	7	30
Trithiozine	18	13	31
Total	41	20	61

- **Formulate the hypothesis testing problem to compare two drugs in terms of differences in proportions - what are the null and alternative?**
- **Which do you think is more appropriate test for such sample sizes - χ^2-test or Fisher's exact test?**
- **Perform both tests in R, compare the significance results.**
- **In case of Fisher's exact test - the confidence interval in the output describes which quantity? (HINT - it is not the difference in proportions)**

PROBLEM #3 - 15 points.

Refer back to the data set you used for problem 3 in your HW #1, where you needed to figure out the relationship between two variables (in case you had categorical data there - please find a new data set with two quantitative variables). Now choose one variable as the response, the other as explanatory variable, and perform linear regression in R.

- **Is there a significant relationship between response and expl-ry variable?**
- **If yes - what is the nature of the relationship? E.g. "If variable x increases by 1 unit, then variable y increases/decreases by ..."**
- **Check the diagnostic plots - is the data roughly normal? Are there outliers?**
- **In case there are glaring outliers - try running linear regression with those removed. How do the results look now?**

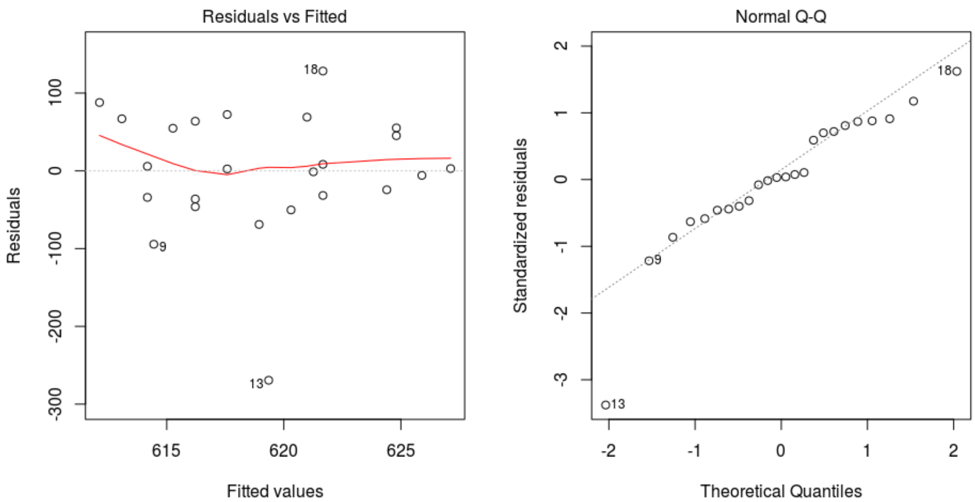
Example: I will check if there is a relationship between students' GPA and their performance on Math GPA.

Code:

```
> library(Stat2Data) # - 'Stat2Data' is one of those packages
> data(SATGPA)       #  where you have to call 'data()' to load the frame.
> attach(SATGPA)
> SAT.lm <- lm(MathSAT ~ GPA,data=SATGPA)
> summary(SAT.lm)
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   576.69     167.20   3.449  0.00229 **
GPA            13.63      53.39   0.255  0.80086
..

> plot(SAT.lm)  ## Observation 13 seems like a clear outlier, let's remove it.
> SAT.lm.new <- lm(MathSAT ~ GPA,data=SATGPA,subset=-13)
...
           Estimate Std. Error t value Pr(>|t|)
(Intercept)   583.23     118.73   4.912 7.37e-05 ***
GPA            15.29      37.91   0.403    0.691
...
```

Plots:



Summary: There seems to be no significant relationship between student's GPA and their performance on Math SAT (very large p-value of 0.8). Upon looking at diagnostic plots, observation 13 appears as if it is an outlier, so we attempt running the regression without it. Result hasn't drastically changed - still no significance claimed (p-value 0.69).

Problem #4: 15 points (+5 EXTRA).

Obtain a multivariate data set containing at least 7 variables (either from your own source, or see `https://vincentarelbundock.github.io/Rdatasets/datasets.html`). DON'T use data sets from *ISwR* package. Select the response variable (as *pemax* was in the case of *cystfib* dataset in class). Then proceed to:

1. **Plot your multivariate dataset. Are there any potential collinearities?**
2. **Perform linear regression of your selected response on all other variables. Any variables demonstrate significance? What about the model overall - is it significant?**
3. **Perform variable selection with "step()" and report the best subset of variables to use. EXTRA 5 POINTS - perform a thorough by-hand variable selection from your domain knowledge considerations (and from data visualization, as in class).**

EXAMPLE (copy-cat from the class): Want to perform linear regression of maximum respiratory pressure on all other variables.

Code:

```
> library(ISwR)
> plot(cystfibr)  # Plot it first.
> cf.lm <- lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc,  # Run lm().
              data=cystfibr)
> summary(cf.lm)
...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582   225.8912   0.779    0.448
age          -2.5420     4.8017  -0.529    0.604
sex          -3.7368    15.4598  -0.242    0.812
height       -0.4463     0.9034  -0.494    0.628
weight        2.9928     2.0080   1.490    0.157
bmp          -1.7449     1.1552  -1.510    0.152
fev1          1.0807     1.0809   1.000    0.333
rv            0.1970     0.1962   1.004    0.331
frc          -0.3084     0.4924  -0.626    0.540
tlc           0.1886     0.4997   0.377    0.711
...
F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195A.
>
> step(cf.lm)  # Variable selection - just provide the final subset by step() here.
```

```
...
Step:  AIC=160.66
pemax ~ weight + bmp + fev1 + rv
        Df  Sum of Sq   RSS    AIC
<none>                 10355 160.66
- rv     1     1183.6 11538 161.36
```

Plot:

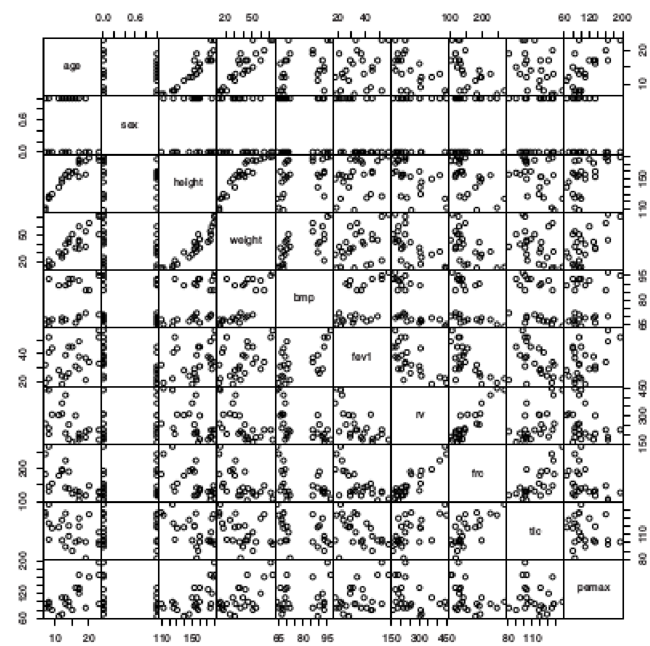


Figure 11.1. Pairwise plots for cystic fibrosis data.

Summary: From plotting the data we can see a clear collinearity between age, height and weight, and also frc and rv variables appear to correlate. The linear regression didn't claim any dominant explanatory variables, but the model overall was deemed significant. The step-wise variable selection procedure revealed $weight, bmp, fev1$ and rv as the most optimal subset to describe $pemax$. (For extra points you'd have to 1) do variable selection by dropping terms according to your logical considerations; 2) explain your logical steps in the summary.)