# MATH 6359, Statistical Computing, Homework 4

Andrey Skripnikov

October 19, 2017

**SUBMISSION GUIDELINES:**

- **Bring the hard-copy of typed up solutions to the class on Tuesday, Oct 24.**
- **Keep it under 5 pages total (all included - text, code, plots, tables). DON'T (!) repeat the problem formulation, go straight to solution.**
- **For problems 1,2 follow the example format in terms of conciseness.**
- **For problems 3,4 - provide both code and your answers in plain English.**
- **Point total is 65 (100%), and on top of that one can get 3 extra credit points total.**

**PROBLEM #1 - 25 points.**

**Find a data set containing two continuous variables $y$ and $x$ (one of them will be chosen by you as a response), and a categorical factor $z$ (make sure to convert it to a factor if it reads as *numeric* at first). As usual, data sets that were covered in class (e.g. *iris*) or from package *ISwR* are NOT allowed. Avoid time series data as well.**

**Having selected the response variable $y$, proceed to:**

- **Formulate the model that only uses factor variable $z$ to explain $y$. Perform simple linear regression of $y$ on $z$, interpret the results - is there a significant effect of the factor on the response? Which factor level corresponds to the REFERENCE group?**
- **Formulate the model that assumes quadratic dependence of $y$ on continuous variable $x$ (don't include factor $z$ into the model). Set the hypotheses to test if quadratic term is significant. Perform quadratic regression of $y$ on $x$, interpret it - does it show potential quadratic relationship between $y$ and $x$?**

- **Formulate the model that assumes dependence of $y$ on both $x$ and $z$, plus it assumes the interaction between $x$ and $z$. Formulate the hypotheses to test if interaction term is significant. Perform full regression of $y$ on both $x$ and $z$ with interaction. Do we witness a significant interaction? If yes - how do you interpret it?**

**EXAMPLE:** I will be looking to analyze the blood pressure (response $y$) in the *bp.obese* data set as a function of obesity (continuous $x$) and gender (factor variable $z$).

**Code (with main raw output included):**

```
> library(ISwR)
> attach(bp.obese)
> sex.fact <- as.factor(bp.obese$sex)                       # Converting 'numeric' to factor.
> levels(sex.fact) <- c("M","F")                            # Making factor names more obvious.
> lm.1 <- lm(bp ~ sex.fact,data=bp.obese)                   # Linear regression of y on z.
> summary(lm.1)                                             # Checking significance.
    ...
    Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  127.955      2.752   46.49   <2e-16 ***
sex.factF     -1.644      3.650   -0.45    0.653
    ...
> lm.2 <- lm(bp ~ obesity + I(obesity^2), data=bp.obese) # Polynomial regression of y on x.
> summary(lm.2)                                             # Checking significance.
    ...
    Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.2404    30.4379   3.162  0.00208 **
obese        23.8192    41.7323   0.571  0.56945
I(obese^2)   -0.2773    13.9658  -0.020  0.98420
    ...
> lm.3 <- lm(bp ~ obesity*sex.fact, data=bp.obese)         # Full interaction model.
> summary(lm.3)                                             # Checking significance of interaction.
    ...
    Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  102.112     18.231   5.601 1.95e-07 ***
obese         21.646     15.118   1.432    0.155
sex.fact          -19.596     21.664  -0.905    0.368
obese:sex.fact      9.558     17.191   0.556    0.579
    ...
```

**Analysis:**

1. Assume the following model:

$$y_i = \beta_0 + \gamma_1 z_i + \epsilon_i,\ \ \epsilon_i \sim N(0, \sigma^2).$$

(if your $z$ has $k > 2$ levels: $y_i = \beta_0 + \gamma_1 z_{1,i} + ... + \gamma_{k-1} z_{k-1,i} + \epsilon_i, \; \epsilon_i \sim N(0, \sigma^2)$).

After running linear regression of $y$ on $z$, the smallest $p$-value across the $k-1$ levels was 0.654, pointing to insignificance of sex when explaining blood pressure (if you have $k > 2$ levels and **at least one of those yields a significant $p$-value** $\Longrightarrow$ claim that there **IS** an effect of your factor on the response). **Males** are the reference group.

2. Assume the following model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \; \epsilon_i \sim N(0, \sigma^2)$$

   The hypotheses to test for quadratic term:

$$H_0 : \beta_2 = 0, \; vs \; H_a : \beta_2 \neq 0$$

   Results of $lm()$: we fail to reject $H_0$ due to $p$-value of 0.98 for quadratic term, hence claiming no quadratic relationship between blood pressure and obesity.

3. Assume the following model:

$$y_i = \beta_0 + \beta_1 x_i + \gamma_1 z + \phi_1 x_i z_i + \epsilon_i, \; \epsilon_i \sim N(0, \sigma^2)$$

   (if your $z$ has $k > 2$ levels - see the slide 38 on linear models).

   Hypotheses to test for interaction:

$$H_0 : \phi_1 = 0, \; vs, \; H_a : \phi_1 \neq 0$$

   Results of $lm()$: we fail to reject $H_0$ due to $p$-value of 0.579 for interaction term. **Interpretation (ONLY if you have significance, here I simply do it for demonstration):** $\hat{\phi}_1 = 9.558$ means that **the slope of $y \sim x$ linear relationship for females is** 9.558 **larger than the slope for males**.

**PROBLEM #2 - 20 points.**

**Write your own function performing the ANOVA F-test for multiple group comparisons (details on slides 6 & 7 of ANOVA lectures). As arguments the function has to take:**

1. **a vector $x$ of continuous values**
2. **the corresponding vector of grouping factor values $z$**

**The function has to output:**

1. **FS, the calculated $F$-statistic value.**
2. **p-value.**
3. **Two-element vector of group and residual degrees of freedom.**
4. **Two-element vector of sums of squares for between-group and within-group deviations.**

**Provide two demonstrative calls for factor variables with different numbers of factor levels (see examples). For each of those, also include the comparative call to $anova()$ for the same vectors, make sure the outputs correspond.**

**Example:**

```
> my.F.test <- function(x,z){
  ..That's all you...
  Make sure that your function works
  for factor variable z with arbitrary number of levels
  (HINT - some looping may be involved)
}
>
>
> ## EXAMPLE 1: for comparison between 3 groups ##
>
> x <- runif(n=50,min=100,max=200) # Sampling uniformly from values in-between 100 and 200.
> z.levels <- c("a","b","c")                          # Assume we will have three levels.
> z <- as.factor(sample(z.levels,size=50,replace=T))   # Randomly sample group assignments.
> my.F.test(x,z)                                      # Call for my function.
$FS
[1] 2.360404
$p.val
[1] 0.1054788
```

```
$df
[1]  2 47
$SSD_B
[1] 3009.108
$SSD_W
[1] 29958.45
> anova(lm(x~z))    #  Comparative call for Rs function.
Analysis of Variance Table
Response: x                                        # Pr(>F) = your p.val, F value = FS
          Df  Sum Sq Mean Sq F value Pr(>F)     # Your df should = Dfs here.
z          2  3009.1 1504.55  2.3604 0.1055     # Sum sq z should = your SSD_B.
Residuals 47 29958.5  637.41                    # Sum sq Residuals = your SSD_W.
>
>
> ## EXAMPLE 2: for comparison between 8 groups ##
> x <- runif(n=50,min=100,max=200) # Sampling uniformly from values in-between 100 and 200.
> z.levels <- letters[1:8]                             # Assume we will have 8 levels.
> z <- as.factor(sample(z.levels,size=50,replace=T))   # Randomly sample group assignments.
> my.F.test(x,z)                                       # Call for my function.
$FS
[1] 1.76835
$p.val
[1] 0.1194106
$df
[1]  7 42
$SSD_B
[1] 8813.652
$SSD_W
[1] 29904.67
> anova(lm(x~z))
Analysis of Variance Table
Response: x
          Df  Sum Sq Mean Sq F value Pr(>F)
z          7  8813.7 1259.09  1.7683 0.1194
Residuals 42 29904.7  712.02
```

**PROBLEM #3 - 10 points (provide code + answer questions in plain text).**

**In the *lung* data frame of *ISwR* package, measurements of lung volume (response) are presented for three different methods (groups) applied to certain patients (*subject*).**

- **Conduct a ONE-way ANOVA of lung volume depending on the method - is there significance of *method* variable?**
- **From your own considerations - does it even make much sense to compare the methods without controlling for the patient variable (*subject*)?**
- **Conduct TWO-way ANOVA of lung volume on the *method* and *subject* variables - is it significant now?**

**PROBLEM #4 - 10 points (+3 EXTRA) (provide code + answer questions in plain text).**

**The** *zelazo* **data LIST (notice that it is NOT of class** *data.frame***) from** *ISwR* **package contains the ages of infants (in months) at which they started walking. Infants were separated into four experimental groups - one received active training, second - passive training, third - no training, last one - 8-week controls (+3 EXTRA CREDIT points if you explain to me what it means, because I am yet to figure it out myself). Proceed to:**

- **convert the data into a form suitable for the use of** *lm*() **and** *anova*() **(HINT - you need to create a factor variable vector with groupings)**
- **Conduct the F-test in R. Is there a difference across training groups?**
- **If yes - conduct pairwise testing, accounting for multiplicity of comparisons. If not - still do it, and point out the pair of groups that shows the smallest adjusted** $p$**-value.**