Exercises of Models Course

July 6, 2016

[1]

[1] mailto:

Contents

0.1 ANOVA

0.1.1 Exercise 1

A society that produces ballpoint pens uses a multi-head machine for the production of caps for pens. The productor wants to determine if there are differences on the means of the different heads.

```
data(pencap)
head(pencap)
```

```
## # A tibble: 6 x 2
##   Cavity  Width
##    <int>  <dbl>
## 1      1  9.920
## 2      1 10.006
## 3      1  9.936
## 4      1  9.963
## 5      1 10.004
## 6      1 10.004
```

1. Let us compute descriptive statistics and generate the BW plot of `Width` for each head (`Cavity`).
2. Let us fit a model to check if there are differences between the means of the different groups. Let us use `options(contrasts = c("contr.treatment", "contr.poly"))` contrasts type and comment the results.
3. Let us check model residuals.

0.1.2 Exercise 2

A chemist conducts an experiment to evaluate the efficacy of a solvent to dissolve stains of nail varnish from fabrics. He uses two types of solvent and three types of varnish. The experiment consists of immersing 5 stained fabrics by a certain type of varnish into a bowl with a solvent and of measuring the time (in minutes) necessary to dissolve the stain.

```
data(varnish)
head(varnish)
```

```
## # A tibble: 6 x 3
##   Solvent Varnish  Time
##     <int>   <int> <dbl>
## 1       2       3 32.50
## 2       1       3 30.20
## 3       1       3 27.25
```

```
## 4       2       3 24.25
## 5       2       2 34.42
## 6       2       2 26.00
```

1. Let us compute descriptive statistics and generate the BW plot of `Time` for each level of `Solvent` and `Varnish`. Let us draw also a BW plot for each combination of `Solvent` and `Varnish` factor levels.
2. Let us evaluate also if there are differences in solvent types in respect of the varnish types that produced the stain (Considerer also the interaction in model formulation).
3. Let us improve the model, dropping non significant terms (one by one).
4. Let us check if the removed effect/s are not significant (models comparison).
5. Let us check (final) models residuals.

0.2 Linear models

0.2.1 Exercise 1

The number of impurities (lumps) present in the containers of paint depends on the rate of agitation applied to the container. A researcher wants to determine the relation between the rate of agitation and the number of lumps, so he conducts an experiment. He applies different rates of agitation (`Stirrate`) to 12 containers of paint and he counts the number of impurities (lumps) present in the containers of paint (`Impurity`).

```
data(paint)
head(paint)
```

```
## # A tibble: 6 x 2
##   Stirrate Impurity
##      <int>    <dbl>
## 1       20      8.4
## 2       38     16.5
## 3       36     16.4
## 4       40     18.9
## 5       42     18.5
## 6       26     10.4
```

1. Let us compute the main descriptive statistics and generate the BW plot of `Impurity`. Let us compute also the correlation and the correlation plot between `Impurity` and `Stirrate`.
2. Let us graphically represent the relation between these variables (add regression line to the graph).
3. Let us compute a simple linear regression between `Impurity` and `Stirrate`. Does `Stirrate` influence `Impurity`? How?
4. Let us check (final) models residuals.

0.2.2 Exercise 2

A pressure switch has a membrane whose thickness (in mm) influences the pressure required to trigger the switch itself. The aim is to determine the thickness of the membrane for which the switch "trig" with a pressure equal to 165 ± 15 KPa. 25 switches with different thickness (`DThickness`) of the membrane was analysed, measuring the the pressure at which each switch opens (KPa) (`SetPoint`).

```
data(switcht)
head(switcht)
```

```
## # A tibble: 6 x 2
##   DThickness SetPoint
##        <dbl>    <dbl>
## 1        0.9  223.523
## 2        0.6  157.131
## 3        0.5  149.307
## 4        0.8  200.146
## 5        0.8  199.974
## 6        0.7  166.919
```

1. Let us compute the descriptive statistics and generate the BW-plot of the variable `SetPoint`.
2. Let us graphically represent the relation between `DThickness` and `SetPoint`(add regression line to the graph).
3. Let us compute a linear regression between `DThickness` and `SetPoint` and check the residuals of the fitted model. Does `DThickness` influences `SetPoint`? Is the model correct?
4. Let us improve the model and compare it with the previous model, if necessary.
5. Let us produce the graph of predicted values from the best model estimated.

0.2.3 Exercise 3

The engeneers want to reduce the knocking of the engines. Before doing this, they have to identify which variables influence this phenomenon. Data are randomly collected from 13 engines and contains the following variables:

- `Spark`: indicates the time of advance of the spark plug ignition;
- `AFR`: indicates the air fuel ratio (Air Fuel Ratio);
- `Intake`: indicates the inlet temperature;
- `Exhaust`: indicates the exhaust temperature;
- `Knock`: indicates the knocking of the engine.

```
data(knock)
head(knock)
```

```
## # A tibble: 6 x 5
##   Spark   AFR Intake Exhaust Knock
##   <dbl> <dbl>  <int>   <int> <dbl>
## 1  14.5  13.9     31     638  83.7
## 2  14.4  13.8     32     643  84.0
## 3  13.3  13.7     35     629  84.1
## 4  12.7  13.8     31     669  84.2
## 5  13.3  13.9     30     697  84.4
## 6  13.4  15.2     31     700  88.4
```

1. Let us produce a matrix of scatterplots and comment the results.
2. Let us compute a multiple linear regression between `Knock` and the predictors (additive model).
3. Is It possible to improve the resulting model? How?
4. Compare the initial with the final model.

0.2.4 Exercise 4

A researcher wants to determine which variables are related with the percentage of mortality. Data contains the following variables:

- `Rain`: indicates the annual average rainfall;
- `JanTemp`: indicates the average temperatures in January;
- `JulyTemp`: indicates the average temperatures in July;
- `PctOver65`: indicates the percentage of the population over 65 years;
- `HHSize`: indicates the average size of housing;
- `Education`: indicates the years of education;
- `PctHomesLiveable`: indicates the percentage of "habitable" homes;
- `PopDensity`: indicates the density of population;
- `PctLowIncome`: indicates the percentage of low-income families;
- `PctWhiteCollar`: indicates the percentage of employees;
- `Hydrocarbon`: indicates the pollution level by hydrocarbons;
- `NititeOxide`: indicates the pollution level of nitrite oxide;
- `SulphurDioxide`: indicates the pollution level of sulfur dioxide;

- `RelHum`: indicates the annual average relative humidity at 1 PM;
- `MortalityRate`: indicates the mortality rate for 100000 people.

```
data(mortality)
head(mortality)
```

```
## # A tibble: 6 x 15
##    Rain JanTemp JulyTemp PctOver65 HHSize Education PctHomesLiveable
##   <int>   <int>    <int>     <dbl>  <dbl>     <dbl>            <dbl>
## 1    36      27       71       8.1   3.34      11.4             81.5
## 2    35      23       72      11.1   3.14      11.0             78.8
## 3    44      29       74      10.4   3.21       9.8             81.6
## 4    47      45       79       6.5   3.41      11.1             77.5
## 5    43      35       77       7.6   3.44       9.6             84.6
## 6    53      45       80       7.7   3.45      10.2             66.8
## # ... with 8 more variables: PopDensity <int>, PctLowIncome <dbl>,
## #   PctWhiteCollar <dbl>, Hydrocarbon <int>, NititeOxide <int>,
## #   SulphurDioxide <int>, RelHum <int>, MortalityRate <dbl>
```

1. Let us build the complete model, with all the regressors (additive model).
2. Let us reach to a model in which all the regressors have significant terms at the 5% level, by eliminating a regressor at a time.
3. Let us check the residuals of the final model.
4. Let us compare the initial and the final model.

0.2.5 Exercise 5

A researcher wants to establish if there is a relation between the weight of the body and the weight of the brain of fifteen mammal (african elephant, cow, monkey, man, gray wolf, red fox, armadillo, chinchilla and so on).

```
data(brainbod)
head(brainbod)
```

```
## # A tibble: 6 x 3
##    Species    Body  Brain
##     <fctr>   <dbl>  <dbl>
## 1  afeleph 6654.00 5712.0
## 2      cow  465.00  423.0
## 3   donkey  187.00  419.0
## 4      man   62.00 1320.0
## 5 graywolf   36.33  119.5
## 6   redfox    4.24   50.4
```

1. Let us show the descriptive graphics (histogram and BW plot) of the `Body` and `Brain` variables.
2. Let us show a scatterplot of the two variables. Let us estimate a linear regression model and check the model residuals. The linear model between the two variables is adequate? Why?
3. How could the model be improved? What variable transformation can be applied? Generate BW plot, histogram and scatterplot of the transformed variables.
4. Let us estimate a linear regression model with the transformed variables and check the model residuals.

0.3 Generalized Linear models

0.3.1 Exercise 1

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

```
data(admission)
head(admission)
```

```
## # A tibble: 6 x 4
##   admit   gre   gpa  rank
##   <int> <int> <dbl> <int>
## 1     0   380  3.61     3
## 2     1   660  3.67     3
## 3     1   800  4.00     1
## 4     1   640  3.19     4
## 5     0   520  2.93     4
## 6     1   760  3.00     2
```

This dataset has a binary response (outcome, dependent) variable called `admit`. There are three predictor variables: `gre`, `gpa` and `rank`. We will treat the variables `gre` and `gpa` as continuous. The variable `rank` takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest.

1. Let us compute the descriptive statistics for all variables. Let us produce the plots showing the relation between `admit` and `GPA` by `rank`, with lowess lines and the plots showing the relation between `admit` and `GRE` by `rank`, with lowess lines.
2. Let us estimate a logistic regression model using the `glm` (generalized linear model) function. First, we convert `rank` to a factor to indicate that `rank` should be treated as a categorical variable.
3. Check the model residuals.
4. Let us compute model predictions.

0.3.2 Exercise 2

We want to analyze if the number of awards earned by students at one high school are related to the type of program in which the student was enrolled (e.g., vocational, general or academic) and to the score on their final exam in math.

```
data(awards)
head(awards)
```

```
## # A tibble: 6 x 4
##      id num_awards  prog  math
##   <int>      <int> <int> <int>
## 1    45          0     3    41
## 2   108          0     1    41
## 3    15          0     3    44
## 4    67          0     3    42
## 5   153          0     3    40
## 6    51          0     1    42
```

```
# Convert prog and id variables as factors
awards <- awards %>% mutate(prog=factor(prog, levels=1:3, labels=c("General", "Academic", "
            id = factor(id))
```

`num_awards` is the outcome variable and indicates the number of awards earned by students at a high school in a year, `math` is a continuous predictor variable and represents students' scores on their math final exam, and `prog` is a categorical predictor variable with three levels indicating the type of program in which the students were enrolled. It is coded as 1 = "General", 2 = "Academic" and 3 = "Vocational".

1. Let us compute some descriptive statistics of all variables and generate an histogram of the number of awards earned by students by program type.
2. Let us perform a Poisson model analysis using the `glm` function, considering the complete (additive) model.
3. Let us try to improve the model removing non significant variable/s. Compares this model with the previous one.
4. Let us compure predictions of the best fitted model.