Experimental Design and Data Analysis: Assignment 5

This assignment consists of 4 exercises. Throughout this assignment tests should be performed using a level of 0.05.

EXERCISE 1
The file `nauseatable` contains data about post-operative nausea after medication against nausea. Two different medicines were administered to patients that complained about post-operative nausea. One of the medicines, Pentobarbital, was administered in two different doses.

1. Set up a `data.frame` in R existing of two columns and 304 rows. One column should contain an indicator whether or not the patient in that row suffered from nausea, and the other column should indicate the medicin. (Use `nausea.frame=data.frame(nausea,medicin)` where `nausea` is a vector 0's and 1's and `medicin` is the vector containing the medicin labels for each patient. Make sure these columns match correctly.)

2. Study the outcome of `xtabs(~medicin+naus)`.

3. Perform a permutation test in order to test whether the different medicins work equally well against nausea. Permute the medicin labels for this purpose. Use as test statistic the chisquare test statistic for contingency tables, which can be extracted from the output of `chisq.test`: `chisq.test(xtabs(~medicin+nausea))[[1]]`.

4. Compare the p-value found by the permutation test with the p-value found from the chisquare test for contingency tables. Explain the difference/equality of the two p-values.

EXERCISE 2
This exercise concerns the data in the file `airpollution`. Investigate which explanatory variables need to be included into a linear regression model with `oxidant` as the response variable. Do this as follows.

1. Make scatter plots of the candidate explanatory variables against each other and against the response variable (see the R-function `pairs()`). Interpret the plots. Do you judge a linear model to be useful here?

2. Determine for each of the explanatory variables the simple linear regression model. Choose the best among these models, and stepwise extend this model by adding one explanatory variable per step on the basis of the determination coefficient. Use a test to investigate whether the extensions are useful. Determine in this way an appropriate linear regression model for these data.

3. Estimate the parameters in the full linear regression model with all explanatory variables in it. Now stepwise decrease this full model with the

aid of tests of the form $H_0 : \beta_i = 0$. Determine in this way an appropriate linear regression model for the data.

4. Present the estimates of the parameters of the final model of your choice.

5. Investigate the normality of the residuals of the chosen model. Do you think, in view of all results, that the chosen linear model is appropriate?

EXERCISE 3

To investigate the influence of the mutation probability on the quality of an evolutionary algorithm for a certain task it is decided to perform a regression analysis with the mutation probability as a numerical variable. The mutation probability is varied from 0.005 to 0.075 with step size 0.005. For each of the 15 values so obtained the algorithm is run 3 times and a measure of quality (small values are good) is obtained. The data of the experiment are given in the file `genal2.txt`.

1. Read in the data and transform them to a `data.frame` object with two columns: the outcome and the mutation probability, in the form:

```
> genal2frame[1:6,]
          y   mut
1 0.5105618 0.005
2 0.3428392 0.005
3 0.5490900 0.005
4 0.4460895 0.010
5 0.5214594 0.010
6 0.4009612 0.010
```

2. In a single plot make 15 boxplots corresponding to the 3 repetitions at each mutation probability. Does the dependence of `y` on `mut` look linear?

3. Fit a multiple linear regression model for `y` on `mut` *and* the square of `mut`. You can do this by first adding the squared mutation probabilities to the data.frame, and next using `lm`, as follows.

```
> genal2frame$mut2=genal2frame$mut^2
> genal2frame[1:6,]
          y   mut    mut2
1 0.5105618 0.005 2.5e-05
2 0.3428392 0.005 2.5e-05
3 0.5490900 0.005 2.5e-05
4 0.4460895 0.010 1.0e-04
5 0.5214594 0.010 1.0e-04
6 0.4009612 0.010 1.0e-04
> genal2lm=lm(y~mut+mut2,data=genal2frame)
```

Describe the model that this fits to the data (in the form $Y = \mu + ?? +$ `error`).

4. Make a plot of $\mu + ??$ as a function of the mutation probability ranging over a grid from 0.004 to 0.08 into 1000 steps. According to this model, what is the optimal mutation probability?

5. How many parameters were estimated from the data? How many would have been estimated from the data if we had performed a one-away analysis of variance with the mutation probabilities defined as labels rather than as numerical variables? How many observations per parameter would that be?

EXERCISE 4

The data in `expensescrime` were obtained to determine factors related to state expenditures on criminal activities (courts, police, etc.) The variables are: `state` (indicating the state in the USA), `expend` (state expenditures on criminal activities in $1000), `bad` (number of persons under criminal supervision), `crime` (crime rate per 100000), `lawyers` (number of lawyers in the state), `employ` (number of persons employed in the state) and `pop` (population of the state in 1000). Perform a regression analysis using `expend` as response variable and `bad`, `crime`, `lawyers`, `employ` and `pop` as independent variables. Present your final model. Your analysis should at least include:

a. investigation of potential or influence points

b. investigation of problems due to collinearity

c. investigation of residuals.

You may use all techniques mentioned on the lecture slides. State clearly all the choices you make during the regression analysis, including arguments for all your choices. (Note that there are several strategies possible!)