MSG500/MVE190
Linear Models - Lab 2

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

November 12, 2012

1 PART I

In this lab you will try your hand at multivariate regression modeling looking at a data set on Nobel Laureates, Chocolate and coffee consumption, and some other variables. Check out the paper in *New England Journal of Medicine*, Oct 12, 2012 by Frank J. Messerli on chocolate consumption and Nobel Laureates. I took that data set and added a few more variables, mostly using information on wikipedia.

Load this data set into R by calling the function `read.table`.

```
> choc <- read.table("chocdata.dat", header = T)
> names(choc)

 [1] "country"        "prizes"         "chocolate"      "coffee"
 [5] "gdp"            "gdponrd"        "life"           "fertility"
 [9] "obesity"        "qualityoflife"  "summerolympic"  "winterolympic"

> dim(choc)

[1] 23 12
```

As you see, the data consists of 23 observations and 12 variables. The number of Nobel Laurates, `prizes`, is the second t column in the data set, the first comprises the countries. Other variables include `chocolate` (consumption in kg per year and capita), `coffee` (same units), `gdp` (gross domestic product per capita), `gdponrd` (percent of gdp spent on research and development, `life` (mean life time), `fertility` (rate), `obesity` (percentage of obese in population), `qualityoflife` (an index that takes life time, economy and health into account), and the cumulate number of gold, silver and bronze medals in the `summerolympic` and `winterolympic` games.

This data set is rather small and so it will be difficult to pick variable transformations and to choose outliers.

1.1 Data exploration

Plot the number of Nobel Laureates versus the other variables (excluding the olympic medals). Can you find a variable transformation that enhances the correlation between the Nobel Laureates and the others?

1.2 Modeling

Is a linear model suitable for this data set (after transformations)? Are any of the coefficients significant? Are there any obvious outliers? Tell me how you identified them - which diagnostic tools were most enlightening?

What happens if you remove the outliers? How much does the fit improve?

1.3 Model selection

Perform backward model selection - comment on the outcome. Does this result agree with the significance analysis of the slope coefficients above? Does the result make sense to you? Any concern about collinearities in this data set?

1.4 Olympic medals

Repeat the above with either summer, winter or total number of olympic medals as your dependent variable. Which variables, if any, seem to be predictive of athletic success?

2 PART II

In the second part of the lab you will try your hand at multivariate regression modeling looking at a data set pertaining to the birthweight of babies. This is a much larger data set than the above and so it is easier to choose appropriate variable transformations and to detect outliers.

Load this data set into R by calling the function `read.table`.

```
> baby <- read.table("Babydata.txt", header = T)
> names(baby)
```

```
[1] "bwt"       "gestation" "parity"    "age"       "height"    "weight"
[7] "smoke"
```

```
> dim(baby)
```

```
[1] 250   7
```

As you see, the data consists of 250 observations and 7 variables. The baby birthweight, `bwt`, is the first column in the data set. Explanatory variables include `gestation` (length of pregnancy in days), `parity` (1 if this is not the first baby), `age` of mother, `height` and `weight` of mother, and finally a variable that indicates if the mother is a `smoker` or not (1 if the mother smokes).

Each of you will work on a different subset of the data for this lab. To achieve this, sample a set of 100 observations:

```
> babynew <- as.data.frame(baby[sample(seq(1, 250), 100), ])
> row.names(babynew) <- seq(1, 100)
```

The above code creates a new data set, `babynew`, which has 100 observations in it. I need to rename the rows 1-100 since R otherwise reports back the original position of each observation in the data set `baby`.

2.1 Data exploration

Examine the data set - are any transformations necessary for a linear model to be sufficient? Report the pairwise correlations between birthweight and the explanatory variables. Comment.

2.2 Modeling

Examine a linear model fit - are there outliers in the data? any other problems? Comment.

Which variables are significant? Interpret the model as much as you feel is possible. Are you concerned about the interpretability? Is there collinearity in the data? Report the R-squared and F-test results - what do these tell you?

2.3 Model selection

Perform backward model selection - comment on the outcome. Does this result agree with the significance analysis of the slope coefficients above?

2.4 Simulations

Here you can choose between two types of simulation studies;
(1) Adding noise to the birthweight data:

```
> babynew2 <- babynew
> babynew2$bwt <- babynew$bwt + rnorm(100, sd = x)
```

Repeat the modeling and selection of variables several times - comment on what outcomes persist across simulated data sets and which do not. (x is a noiselevel you pick.)

(2) Adding a severe collinearity problem to the data:

```
> babynew2 <- babynew
> babynew2$extra <- babynew$age + rnorm(100, sd = x)
```

Here I create an extra variable that is correlated with age. How correlated it is depends on the standard deviation x which you pick yourselves.
Repeat the modeling and selection of variables several times with the extra included and generated each time - comment on what outcomes persist across simulated data sets and which do not.