MSG500/MVE190
Linear Models - Lab 2

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

November 9, 2011

1 Introduction

In this lab you will try your hand at multivariate regression modeling looking at a data set pertaining to the birthweight of babies.

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 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.

3 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?

4 Model selection

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

5 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.