Multiple linear regression: analysis of heart data

January 5, 2016

1 Dataset

Consider predicting the percent heart rate achieved from other basic heart related measurements. We will use the data set from the course web page called `HeartData.txt`. Definitions of the variables included are:

hr:	basal heart rate
bp:	basal blood pressure
pkhr:	peak heart rate
sbp:	systolic blood pressure
mphr:	percentage of maximum predicted heart rate achieved
age:	age
gender:	gender (male = 0, female=1)
baseef:	baseline cardiac ejection fraction (measures heart pumping efficiency)

We will see if we can predict mphr from the other variables using multiple linear regression. First, save the data set on your hard disk, and run the following R program to read the file:

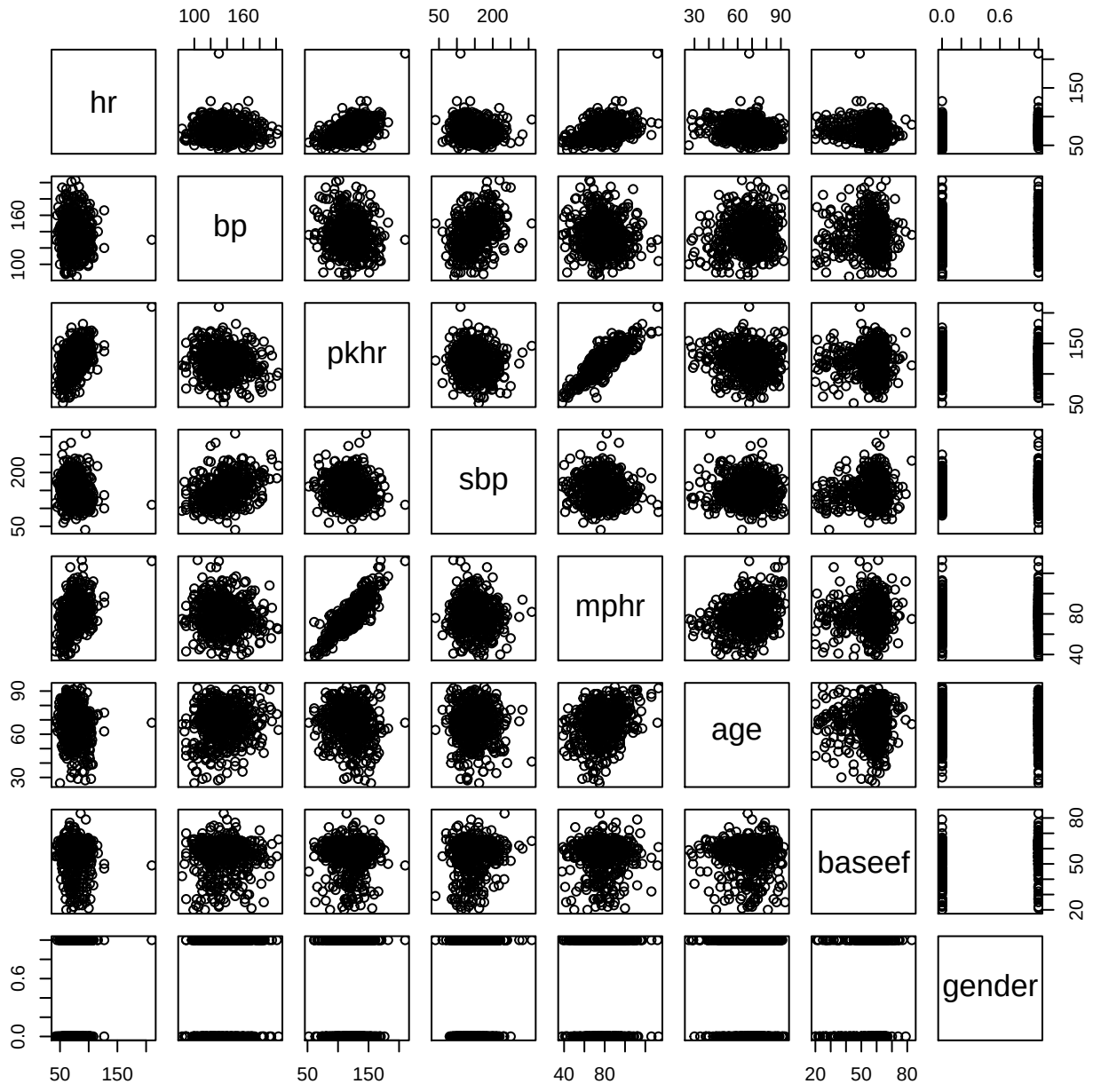
```
heart=read.table(file="HeartData.txt", header=T)
head(heart)
```

```
##   hr  bp pkhr sbp mphr age baseef gender
## 1 92 103  114  86   74  85     27      0
## 2 62 139  120 158   82  73     39      0
## 3 62 139  120 157   82  73     39      0
## 4 93 118  118 105   72  57     42      1
## 5 89 103  129 173   69  34     45      0
## 6 58 100  123 140   83  71     46      0
```

2 Scatterplots of all variables

To verify if the linear assumption is reasonable for this data, is is adviced to make scatterplots of Y versus all continuous explanatory variables. Here is a quick way to produce all scatterplots at once:

```
pairs(heart)
```



Note that some variables look strongly related, e.g., mphr and pkhr (probably not a surprise). Note also that the scatterplot involving the gender variable is meaningless. One can avoid plotting it using the

following command:

```
pairs(heart[,-8])
```

3 Regression model

Now, let's run a multiple linear regression with these variables, using mphr as the dependent variable, and all others as independent variables:

```
regression.out= lm(mphr ~ hr + bp + pkhr + sbp + age + baseef + gender, data=heart)
summary(regression.out)
```

```
##
## Call:
## lm(formula = mphr ~ hr + bp + pkhr + sbp + age + baseef + gender,
##     data = heart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3654  -1.6474   0.2319   1.6702  24.3196
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.712293   2.374200 -14.199  < 2e-16 ***
## hr            0.035350   0.014563   2.427  0.01553 *
## bp           -0.030073   0.010025  -3.000  0.00282 **
## pkhr          0.603127   0.009739  61.928  < 2e-16 ***
## sbp           0.032667   0.005682   5.749 1.49e-08 ***
## age           0.525863   0.016094  32.674  < 2e-16 ***
## baseef        0.010269   0.018996   0.541  0.58902
## gender        0.328743   0.399230   0.823  0.41061
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.405 on 550 degrees of freedom
## Multiple R-squared:  0.9162,Adjusted R-squared:  0.9152
## F-statistic: 859.2 on 7 and 550 DF,  p-value: < 2.2e-16
```

We can get confidence intervals for regression parameters using the `confint` command:

```
confint(regression.out)
```

```
##                      2.5 %       97.5 %
## (Intercept) -38.375903276 -29.04868326
## hr            0.006743777   0.06395708
## bp           -0.049765950  -0.01038055
## pkhr          0.583996404   0.62225759
## sbp           0.021505285   0.04382926
## age           0.494249025   0.55747657
## baseef       -0.027045455   0.04758352
## gender       -0.455459465   1.11294500
```

Take a few minutes to look at each variable, especially at the confidence intervals, and interpret the effect from each variable. Keep in mind the correct way to interpret confidence intervals, according to where the upper and lower interval limits fall in relation to the region of clinical equivalence.

In the above analyses, we simply took all possible variables, and ran a single linear regression model with all variables included. This is, in fact, very bad practice. In general, a better procedure would be something like:

1. Look at various descriptive statistics to get a feel for the data.
2. Separately, graph each independent variable against each dependent variable (sometimes more than one dependent variable is of interest. This allows one to see if trends are perhaps non-linear (possibly leading to data transformations for either X or Y variables), as well as whether the variable looks "important" (but remember that looks can sometimes be deceiving).
3. For all variables being considered, calculate a correlation matrix of each variable against each other variable. This allows one to begin to investigate possible confounding and collinearity.
4. Perform a simple linear regression for each independent variable. This again begins to investigate confounding, as well as providing an initial "unadjusted" view of the importance of each variable, by itself.
5. Think about any "interaction terms" that you may want to try in the model.
6. Perform some sort of model selection technique, or, often much better, think about avoiding any strict model selection by finding a set of models that seem to have something to contribute to overall conclusions.
7. Verify the model's assumptions
8. Based on all work done, draw some inferences and conclusions. Carefully interpret each estimated parameter, perform "model criticism", possibly repeating some of the above steps (for example, run further models), as needed.

9. Other inferences, such as predictions for future observations, and so on.

The above should not be looked at as "rules" never to be broken, but can be used as a rough guide as to how to proceed through a regression analysis. In the next few lectures, we will carefully investigate each of these issues using this same data set and others.