# Math- 521: Lab 3 and Introduction to R (11): Multiple Regression and Model Selection

## 1 Data Descriptions

(1) (Dalgaard, 2002). In chapter 9 of the textbook, Dalgaard presents the analysis related to a study concerning lung function in patients with cystic fibrosis. Data are in the *ISwR* package which can be downloaded from the web. After connecting to the web, all you need to do is to type the following in *R*:

```r
> install.packages("ISwR",.libpaths()[[1]])
```

Then, each time you need to use the data sets included in this package in R, you need to click on "packages" and choose "ISwR" from the packages menu.

Each data set will be available by using the command "data" (see Dalgaard Appendix A and section 9.1).

To read the cystic fibrosis data, I type:

```r
> data(cystfibr)
> cystfibr
```

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>sex</th>
<th>height</th>
<th>weight</th>
<th>bmp</th>
<th>fev1</th>
<th>rv</th>
<th>frc</th>
<th>tlc</th>
<th>pemax</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>0</td>
<td>109</td>
<td>13.1</td>
<td>68</td>
<td>32</td>
<td>183</td>
<td>137</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>1</td>
<td>112</td>
<td>12.9</td>
<td>65</td>
<td>19</td>
<td>449</td>
<td>245</td>
<td>134</td>
<td>85</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>0</td>
<td>124</td>
<td>14.1</td>
<td>64</td>
<td>22</td>
<td>441</td>
<td>268</td>
<td>147</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>1</td>
<td>125</td>
<td>16.2</td>
<td>67</td>
<td>41</td>
<td>234</td>
<td>146</td>
<td>124</td>
<td>85</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>0</td>
<td>127</td>
<td>21.5</td>
<td>93</td>
<td>52</td>
<td>202</td>
<td>131</td>
<td>104</td>
<td>95</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>0</td>
<td>130</td>
<td>17.5</td>
<td>68</td>
<td>44</td>
<td>308</td>
<td>155</td>
<td>118</td>
<td>80</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>1</td>
<td>139</td>
<td>30.7</td>
<td>89</td>
<td>28</td>
<td>305</td>
<td>179</td>
<td>119</td>
<td>65</td>
</tr>
<tr>
<td>8</td>
<td>12</td>
<td>1</td>
<td>150</td>
<td>28.4</td>
<td>69</td>
<td>18</td>
<td>369</td>
<td>198</td>
<td>103</td>
<td>110</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>0</td>
<td>146</td>
<td>25.1</td>
<td>67</td>
<td>24</td>
<td>312</td>
<td>194</td>
<td>128</td>
<td>70</td>
</tr>
<tr>
<td>10</td>
<td>13</td>
<td>1</td>
<td>155</td>
<td>31.5</td>
<td>68</td>
<td>23</td>
<td>413</td>
<td>225</td>
<td>136</td>
<td>95</td>
</tr>
<tr>
<td>11</td>
<td>13</td>
<td>0</td>
<td>156</td>
<td>39.9</td>
<td>89</td>
<td>39</td>
<td>206</td>
<td>142</td>
<td>95</td>
<td>110</td>
</tr>
<tr>
<td>12</td>
<td>14</td>
<td>1</td>
<td>153</td>
<td>42.1</td>
<td>90</td>
<td>26</td>
<td>253</td>
<td>191</td>
<td>121</td>
<td>90</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>0</td>
<td>160</td>
<td>45.6</td>
<td>93</td>
<td>45</td>
<td>174</td>
<td>139</td>
<td>108</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>1</td>
<td>158</td>
<td>51.2</td>
<td>93</td>
<td>45</td>
<td>158</td>
<td>124</td>
<td>90</td>
<td>80</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>1</td>
<td>160</td>
<td>35.9</td>
<td>66</td>
<td>31</td>
<td>302</td>
<td>133</td>
<td>101</td>
<td>134</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>1</td>
<td>153</td>
<td>34.8</td>
<td>70</td>
<td>29</td>
<td>204</td>
<td>118</td>
<td>120</td>
<td>134</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>0</td>
<td>174</td>
<td>44.7</td>
<td>70</td>
<td>49</td>
<td>187</td>
<td>104</td>
<td>103</td>
<td>165</td>
</tr>
<tr>
<td>18</td>
<td>17</td>
<td>1</td>
<td>176</td>
<td>60.1</td>
<td>92</td>
<td>29</td>
<td>188</td>
<td>129</td>
<td>130</td>
<td>120</td>
</tr>
<tr>
<td>19</td>
<td>17</td>
<td>0</td>
<td>171</td>
<td>42.6</td>
<td>69</td>
<td>38</td>
<td>172</td>
<td>130</td>
<td>103</td>
<td>130</td>
</tr>
<tr>
<td>20</td>
<td>19</td>
<td>1</td>
<td>156</td>
<td>37.2</td>
<td>72</td>
<td>21</td>
<td>216</td>
<td>119</td>
<td>81</td>
<td>85</td>
</tr>
<tr>
<td>21</td>
<td>19</td>
<td>0</td>
<td>174</td>
<td>54.6</td>
<td>86</td>
<td>37</td>
<td>184</td>
<td>118</td>
<td>101</td>
<td>85</td>
</tr>
<tr>
<td>22</td>
<td>20</td>
<td>0</td>
<td>178</td>
<td>64.0</td>
<td>86</td>
<td>34</td>
<td>225</td>
<td>148</td>
<td>135</td>
<td>160</td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td>0</td>
<td>180</td>
<td>73.8</td>
<td>97</td>
<td>57</td>
<td>171</td>
<td>108</td>
<td>98</td>
<td>165</td>
</tr>
<tr>
<td>24</td>
<td>23</td>
<td>0</td>
<td>175</td>
<td>51.1</td>
<td>71</td>
<td>33</td>
<td>224</td>
<td>131</td>
<td>113</td>
<td>95</td>
</tr>
</tbody>
</table>
The description for the data set is also given at page 228 of the text book (Appendix B): The \textit{cystfibr} data frame has 25 rows and 10 columns. It contains lung function data for for cystic fibrosis patients (7-23 years old).

\textbf{Format:}

This data frame contains the following columns:

- \texttt{age} a numeric vector. Age in years.
- \texttt{sex} a numeric vector code. 0: male, 1: female.
- \texttt{height} a numeric vector. Height (cm).
- \texttt{weight} a numeric vector. Weight (kg).
- \texttt{bmp} a numeric vector. Body mass (% of normal).
- \texttt{fev1} a numeric vector. Forced expiratory volume.
- \texttt{rv} a numeric vector. Residual volume.
- \texttt{frc} a numeric vector. Functional residual capacity.
- \texttt{tlc} a numeric vector. Total lung capacity.
- \texttt{pemax} a numeric vector. Maximum expiratory pressure.

2 Preliminary Exploratory Analysis

First of all, by typing:

\begin{verbatim}
> plot(cystfibr)
\end{verbatim}

you can obtain a matrix for all pairwise scatterplots associated with the data set (figure 1). This is an extremely powerful tool in visualizing the multivariate data.

secondly, by typing:

\begin{verbatim}
> attach(cystfibr)
\end{verbatim}

the default data set of in R is going to be \textit{cystfibr}. This is really important because to see the column labeled as height instead of typing \begin{verbatim}> cystfibr[,"height"]\end{verbatim} I only need to type: \begin{verbatim}> height.\end{verbatim}
figure 1. This grid of scatterplots reflects the pairwise relationships between all variables of interest.
Also, you can get a matrix for all the pairwise correlations by typing:

```
> cor(cystfibr)
```

```
     age  sex height weight bmp fev1
age 1.000000 0.2713516 -0.6215056 -0.5951995 -0.5695199 0.2713516
sex -0.2713516 1.0000000 -0.6242769 -0.6242769 -0.6242769 -0.6242769
height -0.6215056 -0.6242769 1.0000000 -0.6215056 -0.6215056 -0.6215056
weight -0.5951995 -0.6242769 -0.6215056 1.0000000 -0.6215056 -0.6215056
bmp -0.5695199 -0.6242769 -0.6215056 -0.6215056 1.0000000 -0.5695199
fev1 -0.2713516 -0.6242769 -0.6215056 -0.6215056 -0.5695199 1.0000000
```

note that this will enable us to browse through all the pairwise linear associations.

### 3 Multiple Regression Modeling

Here, the response variable is `pemax`. All the other variables are considered as the independent variables.

One might be interested to start with the most complete model. That is, we consider the additive model of:

\[
pemax = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{sex} + \beta_3 \times \text{height} + \beta_4 \times \text{weight} + \beta_5 \times \text{bmp} + \beta_6 \times \text{fev1} + \beta_7 \times \text{rv} + \beta_8 \times \text{frc} + \beta_9 \times \text{tlc}
\]  

(1)
here is the output for the full model:

```r
> summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc))
```

**Call:**

```r
lm(formula = pemax ~ age + sex + height + weight + bmp + fev1 +
   rv + frc + tlc)
```

**Residuals:**

```
       Min  1Q Median  3Q Max
-37.338 -11.532  1.081 13.386 33.405
```

**Coefficients:**

```
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582  225.8912   0.779  0.448
   age     -2.5420    4.8017  -0.529  0.604
    sex     -3.7368   15.4598  -0.242  0.812
height     -0.4463    0.9034  -0.494  0.628
   weight    2.9928    2.0080   1.490  0.157
     bmp     -1.7449    1.1552  -1.510  0.152
   fev1      1.0807    1.0809   1.000  0.333
      rv      0.1970    0.1962   1.004  0.331
     frc     -0.3084    0.4924  -0.626  0.540
tlc        0.1886    0.4997   0.377  0.711
```

Residual standard error: 25.47 on 15 degrees of freedom
Multiple R-Squared: 0.6373, Adjusted R-squared: 0.4197
F-statistic: 2.929 on 9 and 15 DF, p-value: 0.03195

```r
> anova(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc))
```

**Analysis of Variance Table**

```
Response: pemax
    Df  Sum Sq Mean Sq  F value  Pr(>F)
age    1 10098.5 10098.5 15.5661 0.001296 **
sex    1   955.4   955.4  1.4727 0.243680
height 1  155.0  155.00  0.2389 0.632089
weight 1  632.3  632.30  0.9747 0.339170
bmp    1 2862.2 2862.22 4.4119 0.053010 .
fev1   1  1549.1 1549.10 2.3878 0.143120
rv     1  561.9  561.90  0.8662 0.366757
frc    1  194.6  194.60  0.2999 0.592007
tlc    1   92.4   92.40  0.1424 0.711160
Residuals 15 9731.2   648.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```
Analyze the output.

To see how the F-tests for the `anova` command are calculated you can run:

```r
> model.full<-lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
> model.age<-lm(pemax~age)
> anova(model.full,model.age)
```

Analysis of Variance Table

| Model 1: pemax ~ age + sex + height + weight + bmp + fev1 + rv + frc + tlc |
| Model 2: pemax ~ age |
| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
| 1       | 15  | 9731.2 |
| 2       | 23  | 16734.2 | -8 | -7002.9 | 1.3493 | 0.2936 |

This confirms the `anova` output and it basically reflects the scenario in which we would single out "age" against all the other variables. The interpretation is that by collapsing all of the other variables, there will be a smaller ANOVA table in which we would have one variable of interest "age", with the F statistic obtained by dividing $M_{age}$ by MSE form the full model.
3.1 Steps in Dalgaard

• Analyze the following steps in search of significant variables.

```r
model.no.tlc <- lm(pemax ~ age + sex + height + weight + bmp + fev1 + rv + frc)
summary(model.no.tlc)
```

```r
call:
lm(formula = pemax ~ age + sex + height + weight + bmp + fev1 + 
    rv + frc)

residuals:
    Min     1Q   Median     3Q    Max
-38.0718 -10.0665  0.1132  13.5259  36.9902

coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               221.8055   185.4350   1.196  0.2491
age                      -3.1346    4.4144   -0.710  0.4879
sex                      -4.6933   14.8363   -0.316  0.7558
height                   -0.5428    0.8428   -0.644  0.5286
weight                    3.3157    1.7672    1.876  0.0790 .
bmp                      -1.9403    1.0047   -1.931  0.0714 .
fev1                      1.0183    1.0392    0.980  0.3417
rv                        0.1857    0.1887    0.984  0.3396
frc                       -0.2605    0.4628   -0.563  0.5813

---
Multiple R-Squared: 0.6339, Adjusted R-squared: 0.4508
```

```r
model.no.tlc.frc <- lm(pemax ~ age + sex + height + weight + bmp + fev1 + rv)
summary(model.no.tlc.frc)
```

```r
call:
lm(formula = pemax ~ age + sex + height + weight + bmp + fev1 + 
    rv)

residuals:
    Min     1Q Median     3Q    Max

coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              166.71822  154.31294   1.080  0.2951
age                       -1.81783   3.66773   -0.496  0.6265
```

7
|    | Estimate | Std. Error | t value | Pr(>|t|) |
|----|----------|------------|---------|----------|
| (Intercept) | 260.6313 | 120.5215 | 2.163   | 0.0443 * |
| age | -2.9062 | 3.4898 | -0.833  | 0.4159   |
| sex | -1.2115 | 11.8083 | -0.103  | 0.9194   |
| height | -0.6067 | 0.7655 | -0.793  | 0.4384   |
| weight | 3.3463 | 1.4719 | 2.273   | 0.0355 * |
| bmp | -2.3042 | 0.9136 | -2.522  | 0.0123 * |
| fev1 | 1.0274 | 0.6329 | 1.623   | 0.1219   |

> model.no.tlc.frc.rv.fev1<-lm(pemax~age+sex+height+weight+bmp)
> summary(model.no.tlc.frc.rv.fev1)

Call:
`lm(formula = pemax ~ age + sex + height + weight + bmp)`

Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>-43.194</td>
<td>-9.412</td>
<td>9.157</td>
<td>40.112</td>
<td></td>
</tr>
<tr>
<td>1Q</td>
<td>-2.425</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>9.157</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3Q</td>
<td>40.112</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients:
> model.no.tlc.frc.rv.fev1.sex<-lm(pemax~age+height+weight+bmp)
> summary(model.no.tlc.frc.rv.fev1.sex)

Call:
  lm(formula = pemax ~ age + height + weight + bmp)

Residuals:
   Min     1Q Median     3Q    Max
-41.501 -15.460 -2.838  11.082  42.991

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  274.5307   125.5745   2.186 0.040906 *
age         -3.0832    3.65659 -0.843 0.409107
height      -0.6985    0.80082 -0.872 0.393354
weight      3.6338    1.53542  2.367 0.028217 *
bmp         -1.9621    0.93167 -2.106 0.047955 *

---
Multiple R-Squared: 0.5186,   Adjusted R-squared: 0.4223

> model.no.tlc.frc.rv.fev1.sex.age<-lm(pemax~height+weight+bmp)
> summary(model.no.tlc.frc.rv.fev1.sex.age)

Call:
  lm(formula = pemax ~ height + weight + bmp)

Residuals:
   Min     1Q Median     3Q    Max
-41.794 -11.764 -1.218  13.202  43.631

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 245.3936    119.8927   2.047 0.053411 .
  height      3.5546     0.7962  4.431 0.000286 ***
  weight      3.6338     1.5354  2.367 0.028217 *
  bmp         -1.9621     0.9317 -2.106 0.047955 *

---
Multiple R-Squared: 0.5480,   Adjusted R-squared: 0.429
> model.no.tlc.frc.rv.fev1.sex.age.height<-lm(pemax~weight+bmp)
> summary(model.no.tlc.frc.rv.fev1.sex.age.height)

Call:
  lm(formula = pemax ~ weight + bmp)

Residuals:

    Min     1Q   Median     3Q    Max
-42.924 -13.399    4.361  16.642  48.404

Coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  124.8297    37.4786  3.331  0.003032 **
weight       1.6403     0.3900  4.206  0.000365 ***
bmp          -1.0054     0.5814 -1.729  0.097797 .

---

Multiple R-Squared: 0.4749,   Adjusted R-squared: 0.4271

> model.no.tlc.frc.rv.fev1.sex.age.height.bmp<-

> summary(model.no.tlc.frc.rv.fev1.sex.age.height.bmp)

Call:
  lm(formula = pemax ~ weight)

Residuals:

    Min     1Q   Median     3Q    Max
-44.305 -22.691    2.230  15.908  48.408

Coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.5456    12.7016  5.003 4.63e-05 ***
weight        1.1867     0.3009  3.944 0.000646 ***

---

Multiple R-Squared: 0.4035,   Adjusted R-squared: 0.3776

10