

Topic 11

Multiple regression

Sergey Mastitsky ©

Klaipeda, 28-30 September 2011

Multiple regression

- We'll consider regression models with multiple predictors:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

- Model specification is similar to that of the simple regression analysis and ANOVA
- The new part is the **model search**, i.e. selecting a subset of predictors that describe the response variable sufficiently well

11. Multiple regression

11.1. Plotting multivariate data

For your personal use only.
Public presentation not allowed

An example: cystic fibrosis data

```
> library(ISwR)
> data(cystfibr); attach(cystfibr)
> head(cystfibr)
> help("cystfibr")
```

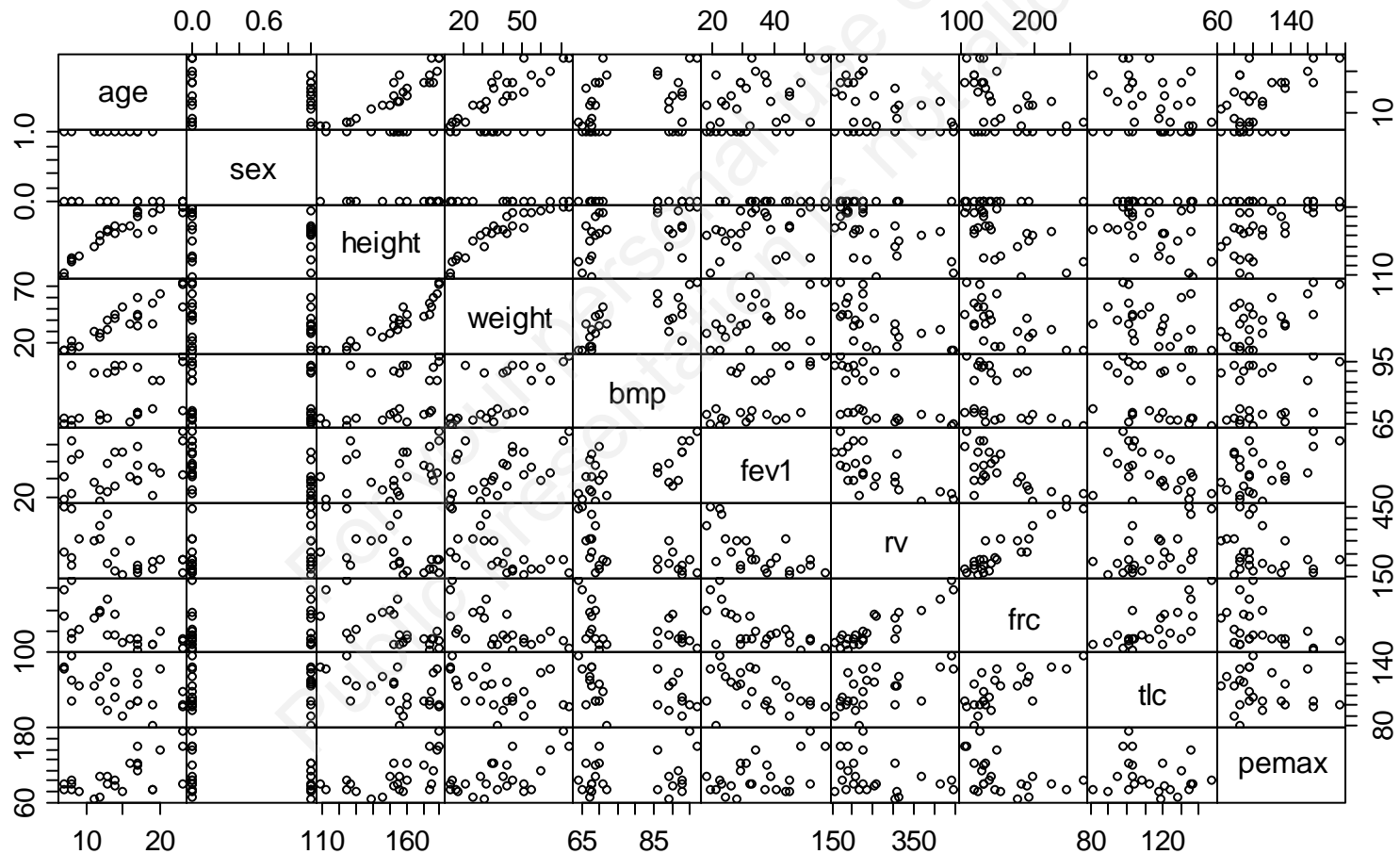
Predictors

Response: *maximum expiratory pressure*

	age	sex	height	weight	bmp	fev1	rv	frc	tlc	pemax
1	7	0	109	13.1	68	32	258	183	137	95
2	7	1	112	12.9	65	19	449	245	134	85
3	8	0	124	14.1	64	22	441	268	147	100
4	8	1	125	16.2	67	41	234	146	124	85
5	8	0	127	21.5	93	52	202	131	104	95
6	9	0	130	17.5	68	44	308	155	118	80

Graphical EDA of multivariate data

```
> pairs(cystfibr, gap = 0, cex.lab = 0.9)
```



11. Multiple regression

11.2. Model specification and output

For your personal use only.
Public presentation not allowed

Model specification in R

- A multiple regression analysis is done by setting up a model formula with “+” between the predictor variables
- Although some predictors are not likely to be correlated with `pemax`, we will initially include all of them into the model
- Such a model is called **saturated**:

```
> M0 <- lm(pemax ~ age + sex +  
height + weight + bmp + fev1 + rv  
+ frc + tlc, data = cystfibr)
```

Model output

```
> summary(M0)
```

```
call:
```

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

```
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
```


Zooming into the results...

- Not a single t -test is significant
- Yet, the overall F -test is significant, so there must be an effect somewhere there
- t -tests only say that *no variable must be included* into the model – thus, we can exclude some of them

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

Pay attention to the adjusted R^2

Multiple R-squared: 0.6373, Adjusted R-squared: 0.4197

- Adjusted R^2 is considerably smaller than the multiple R^2
- This is due to the large number of variables relative to the number of degrees of freedom for the variance
- This also suggests reducing the model

The ANOVA table

```
> anova(M0)
```

```
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    155.0   0.2389 0.632089
weight 1   632.3    632.3   0.9747 0.339170
bmp    1  2862.2   2862.2   4.4119 0.053010 .
fev1   1  1549.1   1549.1   2.3878 0.143120
rv     1   561.9    561.9   0.8662 0.366757
frc    1   194.6    194.6   0.2999 0.592007
tlc    1    92.4     92.4   0.1424 0.711160
Residuals 15  9731.2   648.7
```

The ANOVA table

- Except for t_{lc} , there is practically no correspondence between F-test and t-tests, in particular `age` is now significant
- That is because **F-tests are successive**, i.e. they correspond to a stepwise removal (from the bottom upward) of terms from the model until only `age` is left
- Thus, we can remove all the terms except `age`
- (!) But be careful: `age` was left in the model primarily because it was mentioned first in the model specification (more on this issue later on...)

Comparing two models with `anova()`

- We can formally check whether all the other variables, except for `age`, can be removed by fitting two separate models and comparing their RSS with the F -test:

```
> M0 <- lm(pemax ~ age + sex +  
  height + weight + bmp + fev1 + rv  
  + frc + tlc, data = cystfibr)  
> M1 <- lm(pemax ~ age,  
  data = cystfibr)
```

Comparing two models with anova ()


```
> anova(M0, M1)
```

```
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



- RSS (= unexplained variance) in the second model increases (by 7002.9)
- However, this increase is not significant ($P = 0.2936$), suggesting that model 2 is preferable as it has less parameters

Comparing two models with AIC

- The **Akaike Information Criterion** comes from the Information Theory and measures both the goodness of fit and complexity of a model (see ?AIC)

- The smaller the value of AIC, the “better”:

```
> AIC (M0, M1)
> df      AIC
M0  11  242.0525
M1   3  239.6052
```

Nested models

- Always make sure that the models you compare with `anova()` or `AIC()` are **nested**, e.g. one model is a direct reduced version of another one:

```
> M0 <- lm(pemax ~ age + sex +  
  height + weight + bmp + fev1 +  
  rv + frc + tlc, data = cystfibr)  
> M1 <- lm(pemax ~ age,  
  data = cystfibr)
```


11. Multiple regression

11.2. Stepwise backward model search

For your personal use only.
Public presentation not allowed

Stepwise backward model search

- With this approach, usually the most insignificant terms are sequentially removed until all the remaining terms are significant
- In some cases, a certain logical structure can be imposed on the process, e.g. previous knowledge can suggest removing some parameters first

Manual stepwise reduction of the cystic fibrosis model

sex excluded:

```
M3 <- lm(pemax ~ age + height +  
weight + bmp + fev1 + rv + frc +  
tlc, data = cystfibr)  
> summary(M3)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	153.0385	198.7149	0.770	0.452
age	-2.1145	4.3308	-0.488	0.632
height	-0.3948	0.8517	-0.464	0.649
weight	2.8349	1.8420	1.539	0.143
bmp	-1.7416	1.1207	-1.554	0.140
fev1	1.2651	0.7429	1.703	0.108
rv	0.1779	0.1743	1.021	0.323
frc	-0.2483	0.4123	-0.602	0.555
tlc	0.2084	0.4782	0.436	0.669



Manual stepwise reduction of the cystic fibrosis model

tlc excluded:

```
M4 <- lm(pemax ~ age + height +  
weight + bmp + fev1 + rv + frc,  
data = cystfibr)  
> summary(M4)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	198.2942	165.3311	1.199	0.2468
age	-2.6632	4.0438	-0.659	0.5190
height	-0.4896	0.8037	-0.609	0.5505
weight	3.1557	1.6478	1.915	0.0725 .
bmp	-1.9625	0.9753	-2.012	0.0603 .
fev1	1.2479	0.7240	1.724	0.1029
rv	0.1596	0.1651	0.967	0.3472
frc	-0.1765	0.3687	-0.479	0.6384



Manual stepwise reduction of the cystic fibrosis model

frc excluded:

```
M5 <- lm(pemax ~ age + height +  
weight + bmp + fev1 + rv, data =  
cystfibr)  
> summary(M5)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	166.90487	148.47621	1.124	0.2757	
age	-1.81934	3.56030	-0.511	0.6156	
height	-0.41015	0.76930	-0.533	0.6005	
weight	2.87443	1.50613	1.908	0.0724	.
bmp	-1.94908	0.95382	-2.043	0.0559	.
fev1	1.41196	0.62383	2.263	0.0362	*
rv	0.09558	0.09461	1.010	0.3258	

Age is to be removed!



Manual stepwise reduction of the cystic fibrosis model

Final model:

```
M8 <- lm(pemax ~ weight + bmp +  
  fev1, data = cystfibr)
```

```
> summary(M8)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	126.3336	34.7199	3.639	0.001536	**
weight	1.5365	0.3644	4.216	0.000387	***
bmp	-1.4654	0.5793	-2.530	0.019486	*
fev1	1.1086	0.5144	2.155	0.042893	*

The final model still has to be validated by examining the residuals!

Automatic model selection with the `step()` function

- It's possible to perform automatic search of the optimal model, based on the AIC values
- The function `step()` does all the magic:
> `step(M0, direction = "backward")`

Automatic model selection with the `step()` function

Step 1 in automatic model selection:

Start: AIC=169.11


```
pemax ~ age + sex + height + weight + bmp + fev1 + rv + frc +  
      tlc
```

	Df	Sum of Sq	RSS	AIC
- sex	1	37.90	9769.2	167.20
- tlc	1	92.40	9823.7	167.34
- height	1	158.32	9889.6	167.51
- age	1	181.81	9913.1	167.57
- frc	1	254.55	9985.8	167.75
- fev1	1	648.45	10379.7	168.72
- rv	1	653.78	10385.0	168.73
<none>			9731.2	169.11
- weight	1	1441.21	11172.5	170.56
- bmp	1	1480.12	11211.4	170.65

Automatic model selection with the `step()` function

Final iteration in automatic model selection:

Step: AIC=160.66

pemax ~ weight + bmp + fev1 + rv 

	Df	Sum of Sq	RSS	AIC
<none>			10355	160.66
- rv	1	1183.6	11538	161.36
- bmp	1	3072.6	13427	165.15
- fev1	1	3717.1	14072	166.33
- weight	1	10930.2	21285	176.67

Call:

```
lm(formula = pemax ~ weight + bmp + fev1 + rv, data = cystfibr)
```

Coefficients:

(Intercept)	weight	bmp	fev1	rv
63.9467	1.7489	-1.3772	1.5477	0.1257

Checking the model automatically selected by step ()

Checking the automatically selected model:

```
> summary(lm(pemax ~ weight + bmp + fev1 + rv, data = cystfibr))
```

```
> plot(lm(pemax ~ weight + bmp + fev1 + rv, data = cystfibr))
```

Don't fully rely on the automatic procedure.
Use your brain too!

Exercise

- Try

```
> step(M0, direction = "forward")
```

- and

```
> step(M0, direction = "both")
```

For your personal use only.
Public presentation is not allowed