

## Excercise 4

1. Read the file "species.txt". Make a model of biomass as a function of pH. Is this model significantly better than the null model of a single mean?

```
> setwd("~/R-course-2006/lecture9")
> spc=read.table("species.txt",sep="\t",head=T)
> spc
```

	pH	Biomass	Species
1	high	0.46929722	30
2	high	1.73087043	39
3	high	2.08977848	44
4	high	3.92578714	35
5	high	4.36679265	25
6	high	5.48197468	29
7	high	6.68468591	23
8	high	7.51165063	18
9	high	8.13220251	19
10	high	9.57212864	12
11	high	0.08665367	39
12	high	1.23697390	35
13	high	2.53204324	30
14	high	3.40794153	30
15	high	4.60504596	33
16	high	5.36771709	20
17	high	6.56084215	26
18	high	7.24206214	36
19	high	8.50363299	18
20	high	9.39095342	7
21	high	0.76488801	39
22	high	1.17647020	39
23	high	2.32512082	34
24	high	3.22288207	31
25	high	4.13612930	24
26	high	5.13717652	25
27	high	6.42193811	20
28	high	7.06552638	21
29	high	8.74592918	12
30	high	9.98177013	11
31	mid	0.17576270	29
32	mid	1.37677830	30
33	mid	2.55104256	21
34	mid	3.00027434	18
35	mid	4.90562386	13
36	mid	5.34330542	13
37	mid	7.70000000	9
38	mid	0.55368893	24
39	mid	1.99029644	26
40	mid	2.91263671	26
41	mid	3.21645133	20
42	mid	4.97988468	21
43	mid	5.65872290	15

44	mid	8.10000000	8
45	mid	0.73956986	31
46	mid	1.52693420	28
47	mid	2.23212239	18
48	mid	3.88528818	16
49	mid	4.62650541	19
50	mid	5.12096844	20
51	mid	8.30000000	6
52	mid	0.51127858	25
53	mid	1.47823269	23
54	mid	2.93455800	25
55	mid	3.50548891	22
56	mid	4.61790914	15
57	mid	5.69696382	11
58	mid	6.09301688	17
59	mid	0.73006280	24
60	mid	1.15806838	27
61	low	0.10084790	18
62	low	0.13859609	19
63	low	0.86351508	15
64	low	1.29291903	19
65	low	2.46916355	12
66	low	2.36655309	11
67	low	2.62921708	15
68	low	3.25228652	9
69	low	4.41727619	3
70	low	4.78081039	2
71	low	0.05017529	18
72	low	0.48283691	19
73	low	0.65266714	13
74	low	1.55533656	9
75	low	1.67163820	8
76	low	2.87005390	14
77	low	2.51072052	13
78	low	3.49760385	4
79	low	3.67876186	8
80	low	4.83154245	2
81	low	0.28972266	17
82	low	0.07756009	14
83	low	1.42902041	15
84	low	1.12074092	17
85	low	1.50795384	9
86	low	2.32596318	8
87	low	2.99570582	12
88	low	3.53819909	14
89	low	4.36454121	7
90	low	4.87050789	3

```
> l=lm(Biomass~pH,data=spc)
```

```
> anova(l)
```

```
Analysis of Variance Table
```

```
Response: Biomass
```

```

          Df Sum Sq Mean Sq F value    Pr(>F)
pH          2 110.07   55.04  9.8587 0.0001383 ***
Residuals  87 485.67    5.58
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

>

Using bootstrapping, test how robust the result is that the mean for low pH is lower than that for medium, and that is lower than high. How often out of 1000 bootstraps do you get this result? Do the same for median.

```

> dim(spc)
[1] 90 3
> is.lo.hi=function(i) {
  spc=spc[i,]
  T=tapply(spc$Biomass,spc$pH,mean)
  (T["low"]<T["mid"]) & (T["mid"] < T["high"])
}
>
> boots=sapply(1:1000,function(i) sample(1:90,90,rep=T) )
> x=apply(boots,2,is.lo.hi)
> sum(x)
[1] 979
> is.lo.hi.med=function(i) {
  spc=spc[i,]
  T=tapply(spc$Biomass,spc$pH,median)
  (T["low"]<T["mid"]) & (T["mid"] < T["high"])
}
> x=apply(boots,2,is.lo.hi.med)
> sum(x)
[1] 842
>

```

Make a model of biomass vs. species. Notice that species is read in as a number, but that it should actually be a factor. Does species predict biomass?

```

> l=lm(Biomass~Species,data=spc)
> anova(l)
Analysis of Variance Table

Response: Biomass

          Df Sum Sq Mean Sq F value    Pr(>F)
Species    1  59.91   59.91  9.8388 0.002324 **
Residuals 88 535.83    6.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> f.Species=factor(spc$Species)
> l2=lm(Biomass~f.Species,data=spc)

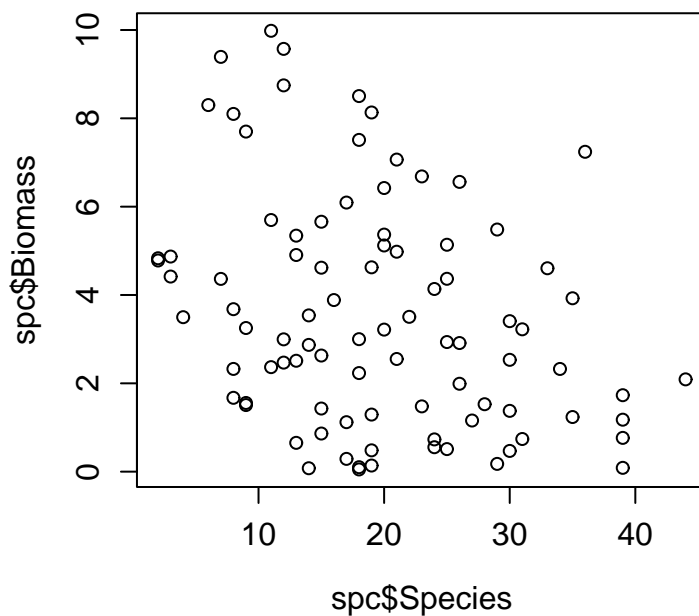
```

```
> anova(12)

Analysis of Variance Table

Response: Biomass
      Df Sum Sq Mean Sq F value Pr(>F)
f.Species 33 202.46    6.14  0.8736 0.6567
Residuals 56 393.28    7.02
```

```
> layout(1)
> plot(spc$Species,spc$Biomass);v()
```



```
>
```

Make a model with species and pH and their interaction, and reduce it to the significant model.

```
> l3=lm(Biomass~f.Species+pH+f.Species:pH,data=spc)
> anova(13)

Analysis of Variance Table

Response: Biomass
      Df Sum Sq Mean Sq F value    Pr(>F)
f.Species  33 202.46    6.14  5.8233 7.852e-07 ***
pH          2 331.59  165.80 157.3701 < 2.2e-16 ***
f.Species:pH 20  25.86    1.29  1.2275  0.2918
Residuals  34  35.82    1.05
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> l4=lm(Biomass~f.Species+pH,data=spc)
```

```
> anova(l3,l4)
```

```
Analysis of Variance Table
```

```
Model 1: Biomass ~ f.Species + pH + f.Species:pH
```

```
Model 2: Biomass ~ f.Species + pH
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	34	35.821				
2	54	61.685	-20	-25.865	1.2275	0.2918

```
> anova(l4)
```

```
Analysis of Variance Table
```

```
Response: Biomass
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
f.Species	33	202.46	6.14	5.3708	2.782e-08 ***
pH	2	331.59	165.80	145.1405	< 2.2e-16 ***
Residuals	54	61.69	1.14		

```
---
```

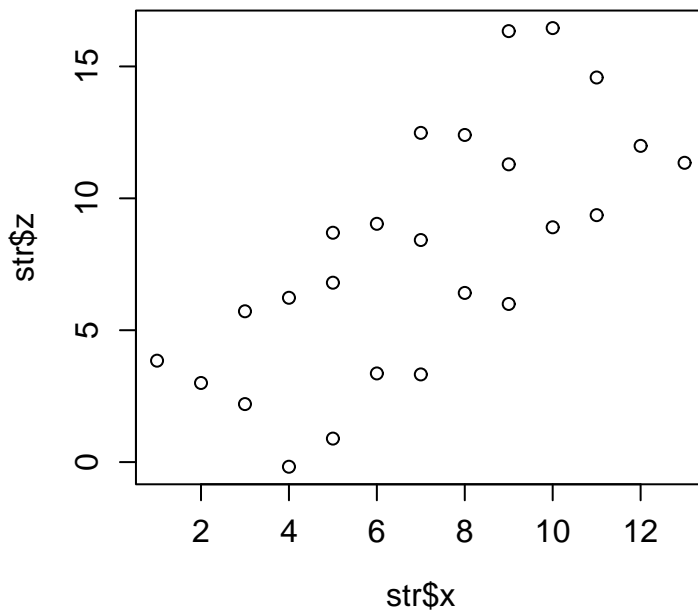
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
>
```

2. Read the file "strange.txt". Plot z vs. x. Do a linear regression of z vs. x. Does z increase or decrease with x? Now do an analysis of variance of linear models for z. In the best model, does z increase or decrease with x?

```
> str=read.table("strange.txt",head=T,sep=",")
```

```
> plot(str$x,str$z);v()
```



```
> l1=lm(z~x,data=str)
> l1
```

```
Call:
lm(formula = z ~ x, data = str)
```

```
Coefficients:
(Intercept)          x
    0.8757         1.0115
```

```
> anova(l1)
```

```
Analysis of Variance Table
```

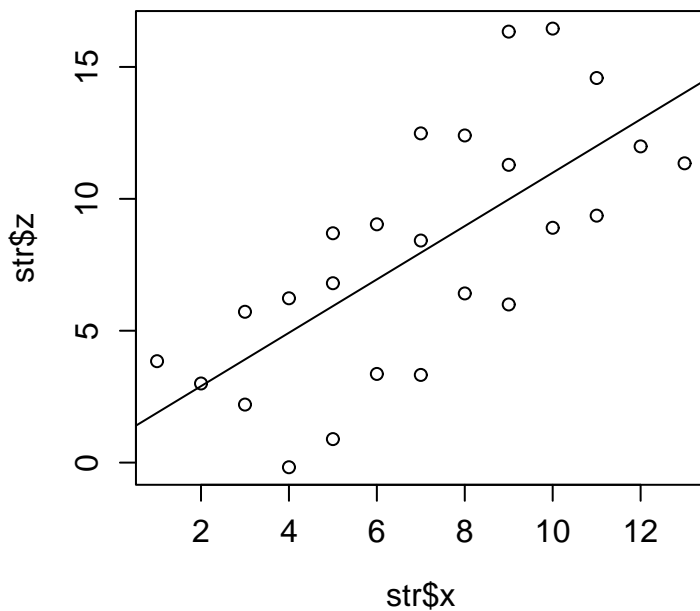
```
Response: z
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	255.761	255.761	22.342	9.198e-05 ***
Residuals	23	263.292	11.447		

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> abline(l1);v()
```



```
> l2=lm(z~x+y+x:y,data=str)
> anova(l2)

Analysis of Variance Table

Response: z

      Df Sum Sq Mean Sq F value    Pr(>F)
x       1 255.761  255.761 343.5759 1.706e-14 ***
y       1 247.596  247.596 332.6085 2.353e-14 ***
x:y     1   0.063   0.063   0.0849  0.7736
Residuals 21  15.633   0.744
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> l3=lm(z~x+y,data=str)
> anova(l2,l3)

Analysis of Variance Table

Model 1: z ~ x + y + x:y
Model 2: z ~ x + y

      Res.Df  RSS Df Sum of Sq    F Pr(>F)
1         21 15.6326
2         22 15.6958 -1   -0.0632 0.0849 0.7736
```

```
> anova(l3)

Analysis of Variance Table

Response: z

      Df Sum Sq Mean Sq F value    Pr(>F)
x       1 255.761  255.761  358.49 4.174e-15 ***
y       1 247.596  247.596  347.04 5.845e-15 ***
Residuals 22  15.696   0.713
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

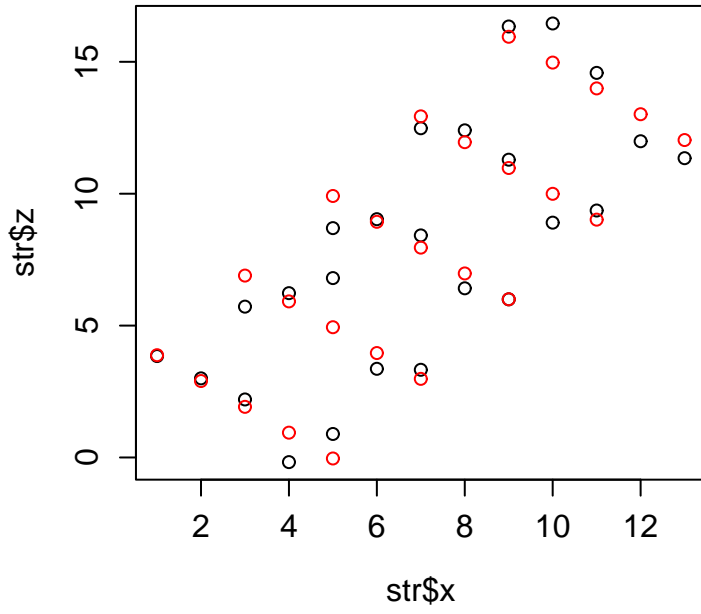
```
> l3

Call:
lm(formula = z ~ x + y, data = str)

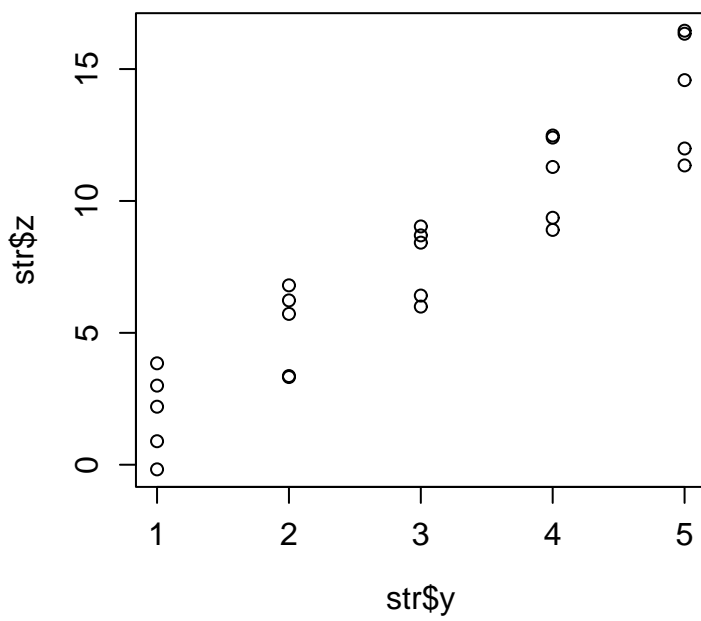
Coefficients:
(Intercept)          x          y
   -0.1195    -0.9789    4.9759
```

```
> plot(str$x,str$z)
> points(str$x,fitted(l3),col=2)
```

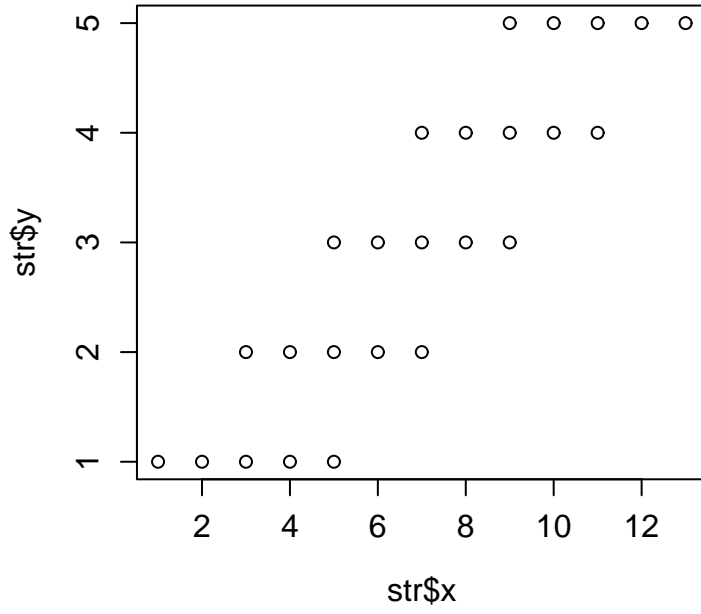
```
> v()
```



```
> plot(str$y, str$z); v()
```



```
> plot(str$x,str$y);v()
```



```
>
```